From the Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

Risk Prediction in Prostate Cancer Diagnostics: Current Challenges and Improvements

Þorgerður Pálsdóttir

Karolinska Institutet

Stockholm 2019

All published papers reproduced with permission
Cover: Watercolor painting by Elín Fanney Guðmundsdóttir
Published by Karolinska Institutet
Printed by E-Print AB 2019

Typeset by the author using LaTeX 2$_\varepsilon$

ISBN 978-91-7831-532-1

Institutionen för Medicinsk Epidemiologi och Biostatistik

Risk Prediction in Prostate Cancer Diagnostics: Current Challenges and Improvements

AKADEMISK AVHANDLING som för avläggande av medicine doktorseksamen vid Karolinska Institutet offentligen försvaras i hörsal Petrén, Nobels väg 12B, Karolinska Institutet, Solna

Fredag 27 september 2019, kl 09.00

By
Þorgerður Pálsdóttir

Principal supervisor:
Associate Professor Martin Eklund
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Co-supervisor:
Doctor Tobias Nordström,M.D.
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Associate Professor Mark Clements
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Doctor Markus Aly, M.D.
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Professor Laufey Tryggvadóttir
University of Iceland
Faculty of Medicine, School of Health Sciences

Opponent:
Professor Donna Ankerst
The Technical University of Munich
Department of Mathematics

Examination board:
Assistant Professor Andreas Josefsson
University of Umeå
Department of Surgical and Perioperative Sciences

Professon Weimin Ye
Karolinska Universitet
Department of Medical Epidemiology and Biostatistics

Senior lecturer Fredrik Granath
Karolinska Universitet
Department of Medicine

For my family Arnaldur, Hilmir Páll, Héðinn and Sigríður Bríet

Abstract

Prostate cancer has been considered a disease of elderly men, and thus historically less focus has been on prostate cancer research than many other cancer types. However, as life expectancy is increasing all over the world, more life years are lost when men are diagnosed with prostate cancer at the age of 70 years now than before. Therefore, it is increasingly important to improve the diagnostic pathway of prostate cancer in modern health care. My thesis aims to address some of the issues in the current prostate cancer diagnostic pipeline using risk prediction models.

Measuring the level of prostate-specific antigen (PSA) in blood is widely used as a blood test to screen for prostate cancer and evidence has shown that mortality decreases with PSA testing. However, because PSA testing has a high false-positive rate, many unnecessary biopsies are performed on healthy men and many men are overdiagnosed with indolent disease (International Society of Urological Pathology (ISUP) grade group 1). In **Study I** the objective was to predict the risk of clinically consequential cancer (ISUP $\geq$ 2) at biopsy and the cumulative probability of having a negative biopsy when being PSA tested with one, two, three, four, or five to eight year intervals. We found that men with a PSA level above 1 ng/mL had an increased risk of ISUP $\geq$ 2 prostate cancer when screened with longer then annual intervals, while men with a PSA level below 1 ng/mL had low risk of ISUP $\geq$ 2 prostate cancer regardless of time between testing. The benefit of a shorter screening interval needs to be balanced with the increased cumulative probability of having a negative biopsy which we found to be twofold for annual vs. biennial testing intervals and threefold for annual vs. triennial testing intervals.

Knowledge about the relationship between PSA, age and different grades of prostate cancer is important for clinicians working with prostate cancer diagnosis because of how widely used the PSA test is. In **Study II** we studied the association between the risk of indolent and clinically consequential prostate cancer (ISUP 1 and ISUP $\geq$ 2) and PSA and age, respectively. Our study cohort comprised of 6.083 biopsied men from the STHLM3 study and 72.996 biopsy cores from those men. In the overall ISUP grade system, lower grades can be masked by higher grades, and thus we studied the associations for both overall ISUP grade and for ISUP grade on each biopsy core. Our results showed that ISUP 1 prostate cancer was not significantly associated with PSA or age, on overall ISUP grade or on individual biopsy core level. In contrast, our results showed that ISUP $\geq$ 2 prostate cancer is significantly associated with increasing PSA level and older age. Our results indicate that PSA leakage of ISUP 1 prostate cancer cells is more similar to that of benign prostate tissue than ISUP $\geq$ 2 prostate cancer tissue.

The use of magnetic resonance imaging (MRI) before biopsy to diagnose prostate cancer has increased in current clinical practice. Combining results from prostate MRI with existing risk prediction models can improve the predictive abilities of the models. The aim of **Study III** was to develop a risk prediction model (S3M-MRI), combining the Stockholm3 score and the

PI-RADS (Prostate Imaging Reporting and data System) score from MRI to predict the risk of ISUP ≥ 2 prostate cancer. We developed the S3M-MRI model using data from the STHLM3-MRI diagnostic study and compared the model performance of the S3M-MRI to the Stockholm3 model and PI-RADS score. We also compared five diagnostic strategies for clinical outcomes. We found that the combined S3M-MRI model had better predictive abilities than both the Stockholm3 and the PI-RADS alone. However, when we compared it to different clinical strategies, the sequential use of the Stockholm3 test followed by MRI on Stockholm3 positive men resulted in similar numbers of performed biopsies and diagnosed ISUP ≥ 2 prostate cancers while saving many MRI scans.

Prostate cancer diagnosis is based on the result of the prostate biopsy and reclassification of ISUP grade on radical prostatectomy samples compared to biopsy is common. In **Study IV** our aim was to study what effect reclassification of disease status based on prostate biopsies has on the performance of prostate cancer risk prediction models using simulations and data from the STHLM3 Radical Prostatectomy Cohort. The cohort comprised of 780 men from the STHLM3 study who were diagnosed with prostate cancer and treated with a radical prostatectomy between 2013 and 2015. We compared four simulated prediction model scenarios with and without error in disease status and calculated the area under the receiver operating characteristics (ROC) curve (AUC) of the Stockholm3 score for predicting clinically significant prostate cancer assessed using biopsy and radical prostatectomy samples. Our simulations showed that fitting a risk prediction model using data with error in the disease status only leads to a small decline in the true predictive performance, but leads to a large decline in apparent predictive performance when evaluated against data with error in the disease status. Moreover, our results showed that the Stockholm3 test has stronger association with clinically significant prostate cancer defined on prostatectomy samples (without errors) than on biopsy samples (with errors).

In conclusion, in this thesis we have aimed to describe a part of the risk associated with diagnosis of prostate cancer as well as developing new prostate cancer risk prediction models. This thesis contributes to the constant pursue of improving the current prostate cancer diagnostic pipeline in order to improve the lives of men screened for or diagnosed with prostate cancer.

List of publications

I. **Thorgerdur Palsdottir**, Tobias Nordström, Andreas Karlsson, Henrik Grönberg, Mark Clements, Martin Eklund
The impact of different prostate specific antigen (PSA) screening intervals on Gleason score at diagnosis and the risk of experiencing false positive biopsy recommendations: A population based cohort study
The BMJ Open 2019

II. **Thorgerdur Palsdottir**, Tobias Nordström, Markus Aly, Johan Lindberg, Mark Clements, Lars Egevad, Henrik Grönberg, Martin Eklund
Are Prostate Specific-Antigen (PSA) and age associated with the risk of ISUP grade 1 prostate cancer? Results from 72 996 individual biopsy cores in 6 083 men from the Stockholm3 study
PLOS ONE 2019

III. **Thorgerdur Palsdottir**, Tobias Nordström, Markus Aly, Fredrik Jäderling, Mark Clements, Henrik Grönberg, Martin Eklund
A Unified Prostate Cancer Risk Prediction Model Combining the Stockholm3 Test and Magnetic Resonance Imaging
European Urology Oncology 2018

IV. **Thorgerdur Palsdottir**, Axel Möller, Markus Aly, Tobias Nordström, Lars Egevad, Henrik Grönberg, Martin Eklund
Implications of ISUP Grade Reclassification Between Biopsy and Radical Prostatectomy Specimens on Prostate Cancer Risk Prediction Models
Manuscript

The articles will be referred to in the text by their Roman numerals, and are reproduced in full at the end of the thesis.

Related publications

- Ola Spjuth, Andreas Karlsson, Mark Clements, Keith Humphreys, Emma Ivansson, Jim Dowling, Martin Eklund, Alexandra Jauhiainen, Kamilla Czene, Henrik Grönberg, Pär Sparén, Fredrik Wiklund, Abbas Cheddad, Thorgerdur Palsdottir, Matthias Rantalainen, Linda Abrahamsson, Erwin Laure, Jan-Eric Litton, Juni Palmgren
  E-Science technologies in a workflow for personalized medicine using cancer screening as a case study
  JAMIA 2017
- Jan Chandra, Markus Aly, Martin Eklund, Thorgerdur Palsdottir, Lars Egevad, Henrik Grönberg, Tobias Nordström
  Lower Urinary Tract Symptoms (LUTS) are not associated with an Increased risk of Prostate Cancer in men 50-69 years with PSA >3 ng/ml
  Submitted 2019
- Axel Möller, Thorgerdur Palsdottir, Martin Eklund, Lars Egevad, Henrik Grönberg, Markus Aly
  Stockholm3 vs. PSA to Predict Clinically Significant Cancer on Radical Prostatectomy
  Manuscript

Contents

List of abbreviations

CI	Confidence Interval
GG	Gleason Grade
IQR	Inter Quartile Range
ISUP	International Society of Urological Pathology
PSA	Prostate-Specific Antigen
OR	Odds Ratio
RP	Radical Prostatectomy
RR	Relative Risk
NB	Net Benefit
DCA	Decision Curve Analysis
MRI	Magnetic Resonance Imaging
ROC	Receiver Operating Characteristics
AUC	Area Under Curve (Receiver Operating Characteristic)
PI-RADS	Prostate Imaging Reporting and Data System
H-L	Hosmer-Lemeshow
RG	Reclassification of ISUP grade

Chapter 1

Introduction

Prostate cancer is the development of cancer in the prostate, a gland in the male reproductive system, and it is highly prevalent in older men. Prostate cancer testing and diagnosis is controversial, and in most developed countries unorganized prostate cancer testing using a blood test called the Prostate-Specific Antigen (PSA) test is very common. The test is an inefficient screening test with many false positive results. The widespread use of the PSA test results in biopsying many healthy men (around half of all biopsies are benign) and overdiagnosis of indolent cancer that can result in overtreatment, infections and anxiety for the men involved. Thus improved methods for prostate cancer testing are in great need in health care systems around the world.

My supervisors have previously developed a prostate cancer risk prediction model called the Stockholm3 test. It was developed in the STHLM3 study, a large cohort study including 59.149 men aged 50–69 years without prostate cancer living in Stockholm, Sweden. The test is based on a risk prediction model using blood based biomarkers, clinical variables and a genetic score to predict the risk of high-grade prostate cancer.

In my thesis I have mainly focused on continuing to improve, validate and further develop the Stockholm3 test, as well as analyzing and describing current PSA testing and prostate cancer risk in Stockholm, Sweden.

Chapter 2

Aims of the thesis

Ineffective screening methods and insufficient risk stratification of screened men are common issues in the current prostate cancer diagnostic pipeline. My thesis aims to describe and address these using risk prediction models.

More specifically, the aims are:

- To study the benefits (decreased risk of higher ISUP grade cancer at diagnosis) and harms (increased risk of having a negative biopsy) of having PSA tests with one, two, three, four or five to eight years testing intervals given a man's current PSA level, age, and family history of prostate cancer.
- To study the associations between the risk of ISUP 1 and ISUP ≥ 2 prostate cancer and PSA and age, respectively.
- To develop a risk prediction model (S3M-MRI) that combines the Stockholm3 test and PI-RADS scores from magnetic resonance imaging (MRI) of the prostate to predict the risk of high-grade prostate cancer.
- To study what effect reclassification of ISUP grade between prostate biopsy and radical prostatectomy specimens has on the predictive performance of prostate cancer risk prediction models using simulations and data from the STHLM3 Radical Prostatectomy Cohort.

Chapter 3

Background

3.1 Prostate Cancer Incidence and Mortality

Prostate cancer is the second most common cancer in the world among men, with an estimated 1.3 million men diagnosed with the disease in 2018. Incidence varies between regions, but it is the most common cancer in men in developed countries [1], where almost 70% of the total number of cases worldwide occur. Highest rates are reported in Australia/New Zealand, North America and Europe, where the use of prostate-specific antigen (PSA) testing and subsequent biopsy has become widespread [2]. In Sweden, the incidence rate of prostate cancer has increased substantially since the introduction of PSA testing for screening of prostate cancer [3]. In 1970, the incidence rate of prostate cancer was 71 new cases per 100,000 men and incidence rose steadily in the following twenty years to 113 new cases per 100,000 men in 1990. After the introduction of PSA testing in 1987, and gradual usage increase during the early 1990s, there was a rapid reported increase in the incidence rate in Sweden. This increase reached a peak in 2004, when 224 new cases were detected per 100,000 men (Figure 3.1). Since 2004, the age-standardized incidence rate has decreased slightly, although prostate cancer was still the most common cancer reported among men in Sweden in 2017 [3, 4].

Among cancers, prostate cancer is the fifth leading cause of death in men in the world, with an estimated 360,000 deaths from prostate cancer in 2018 (around 4% of all cancer deaths in that year). Mortality rates of prostate cancer across developed regions do not differ as widely as incidence rates. However, mortality rates are higher in Southern Africa, the Nordic countries and parts of South America (see Figure 3.2b) [1, 2]. Worldwide, Sweden has among the highest prostate cancer mortality rates (49.5 per 100,000 men in 2014), making prostate cancer the leading cause of death from cancer among Swedish men. Despite the rise in incidence in Sweden, the mortality has decreased since 2005 when it reached a peak of 59 per 100,000 men (Figure 3.1) [3].

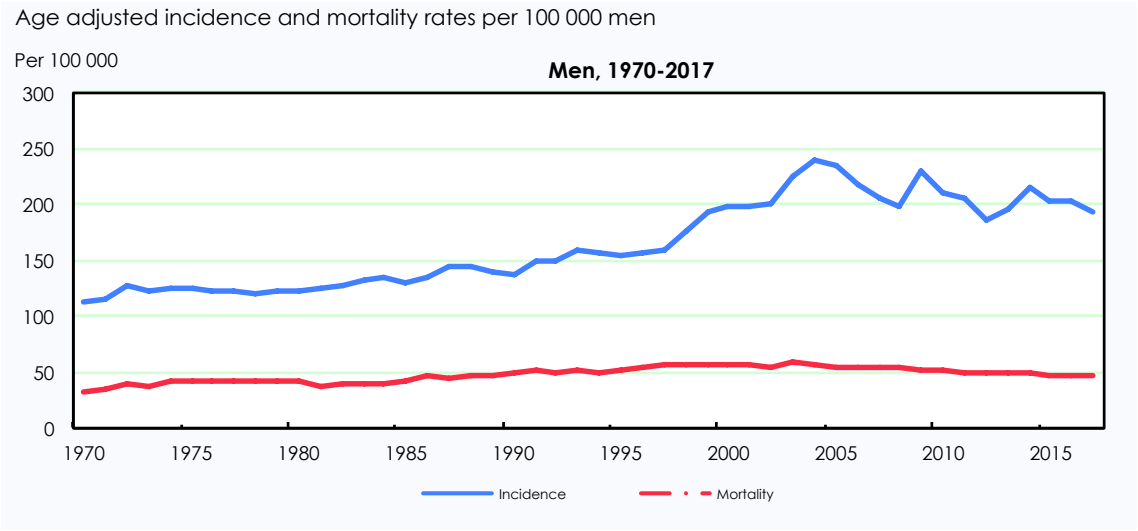


Figure 3.1: Age standardized incidence and mortality rates for prostate cancer in Sweden per 100.000 men. Data from Socialstyrelsen (www.socialstyrelsen.se, accessed 20.02.2019) [3]

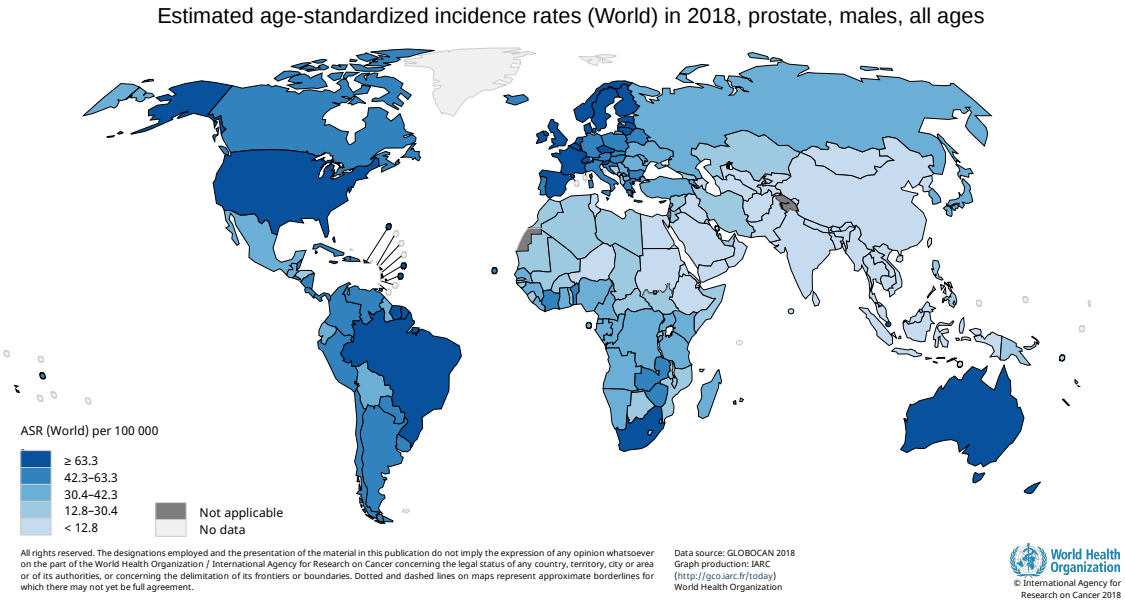
Unlike many other cancer types, prostate cancer incidence and mortality rates are not closely associated. Data from the World Health Organization (WHO) reveals that incidence rate is highest in more developed countries like North America, Australia and Europe (Figure 3.2a), while mortality rate (Figure 3.2b) is highest in Southern Africa. Frequent PSA testing in more developed countries that detect a high rate of low-risk prostate cancers and a higher prostate cancer mortality rate among African-American men are the most plausible reasons for this disparity in incidence and mortality rates around the world.

3.2 Current Prostate Cancer Diagnostic Strategy

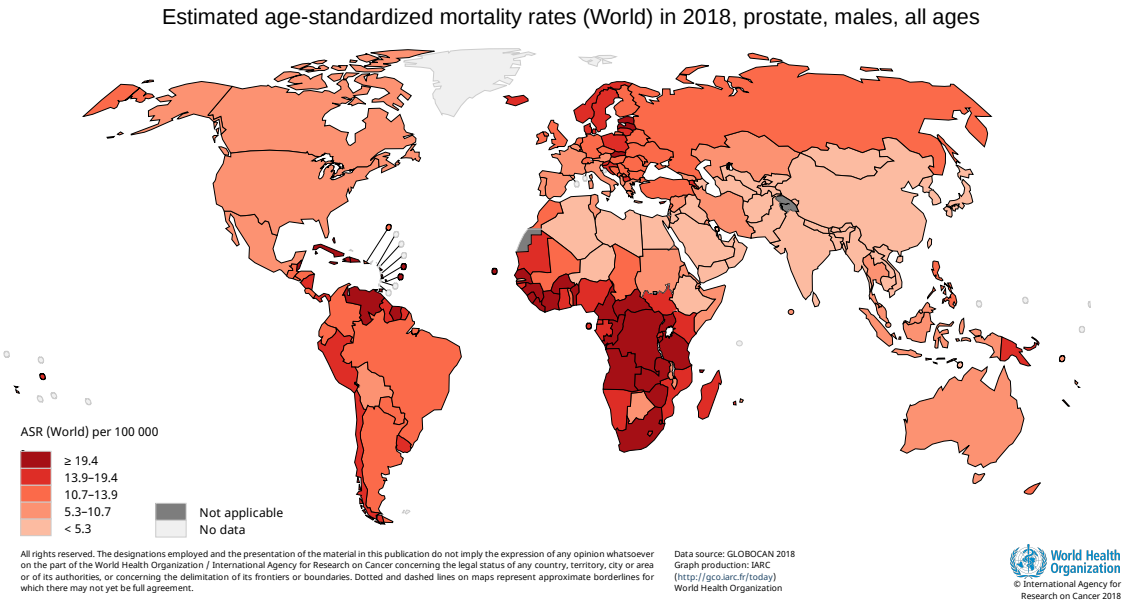
The most common way to diagnose prostate cancer is through assessment of PSA levels, either through routine testing or with suspected increased prostate cancer risk. If the PSA level is elevated, the men are usually recommended to undergo a prostate biopsy, which is subsequently analysed by a pathologist for prostate cancer assessment. After diagnosis of prostate cancer, there are several treatment options available. Figure 3.3 shows an overview of the most common current prostate cancer diagnostic pipeline.

3.2.1 The Prostate-Specific Antigen (PSA) Test and Testing Intervals

As described above, the PSA test is commonly used to test for prostate cancer in developed countries. This development can be explained by greater access to the test, more knowledge among older men of prostate cancer and PSA and better access to



(a) Incidence rates of prostate cancer



(b) Mortality rates of prostate cancer

Figure 3.2: Incidence and mortality rates of prostate cancer in the world in 2018. Data from World Health Organization [1]

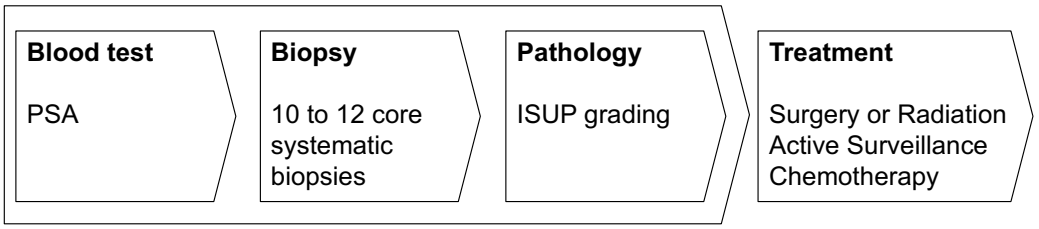


Figure 3.3: The most common prostate cancer diagnosis begins with a PSA blood test, followed by a 10-12 core systematic prostate biopsy. The biopsy sample is then graded by a pathologist and given an ISUP grade and if diagnosed with prostate cancer, the men have an option of different treatments.

health care [5]. Evidence linking PSA testing to reductions in prostate cancer specific mortality are controversial. Many studies have shown that PSA testing leads to reduced prostate cancer specific mortality rates [6, 7, 8], while results from Andriole *et al.* in the PLCO Cancer Screening Trial showed no effect of PSA screening on the rate of death from prostate cancer [9]. However, those results can be at least partly explained with the study design, as a large proportion of the men in the control group were also PSA tested while the study was ongoing and that most men entering the study had already been previously tested. With the high false-positive rate of the PSA test, its wide usage leads to many unnecessary prostate biopsies and overdiagnosis of indolent cancers, resulting in overtreatment [10, 11, 12].

Previous prostate cancer screening trials have used PSA testing intervals of two or four years. However, since these trials were not conducted in a randomized manner with different intervals as a primary objective, little evidence exists supporting an ideal screening interval. Van Leeuwen *et al.* examined screening intervals in a study using data from the European Randomized Study of Screening for Prostate Cancer (ERSPC), and estimated a 43% reduction in the diagnosis of advanced prostate cancer for screening with 2 year intervals (Gothenburg) compared to 4 year intervals (Rotterdam) [13]. In the Gothenburg arm, the incidence of low-grade prostate cancer was 46% higher than in Rotterdam, indicating overdiagnosis of prostate cancer. Modelling studies that have compared screening men annually vs. biennially showed a reduction in the risk of a false-positive test by up to 50% and overdiagnosis by up to 30% while preventing about 80% of deaths prevented with annual screening [14].

Because the ideal PSA screening interval is controversial, the recommendations on testing intervals differ between organizations. The American Cancer Society (ACS) recommendion states that men over 50 years with a PSA below 2.5 ng/mL should be tested biennially, and men with a PSA 2.5 ng/mL or higher should be tested annu-

ally [15]. The American Urological Association (AUA) recommendation is that men aged 55–69 should be tested biennially [16], and the National Comprehensive Cancer Network (NCCN) recommends 2–4 year intervals for men with PSA below 1 ng/mL aged 45–74 and 1–2 year intervals for men with PSA of 1 ng/mL or higher [17]. The European Association of Urology (EAU) recommends PSA testing every 2 years for men at risk, and up to 8 years for men with low risk [18]. Some of these organizations also state in their guidelines that this area of research needs more evidence to find the optimal PSA testing interval. In 2017, the US Preventive Services Task Force (USPSTF) recommended in a draft statement to offer PSA testing to men aged 55 to 69 after individualized decision making – after advocating against PSA testing altogether in 2012 [19, 20]. They give no recommendation for a specific testing interval and mention that more research on the subject is needed.

3.2.2 Diagnosis

Prostate cancer is most commonly diagnosed with biopsies of the prostate, while a small fraction of the men are diagnosed with fine-needle aspiration from the prostate, resected material from transurethral resection of the prostate (TUR-P), or by symptoms. The median age of diagnosis in Sweden has decreased over the years from 74 years in 1996 to 70 years in 2005 (due to the introduction of PSA testing in asymptomatic men) [21]. Below I will briefly describe the *systematic biopsy* procedure – currently the most commonly used biopsy procedure – as well as the variables associated with prostate cancer prognosis, of which the most important is the *Gleason grade*.

Systematic Biopsy

Prostate biopsies are recommended for men with an increased risk of prostate cancer due to relevant risk factors, which include elevated PSA levels, abnormal digital rectal exam, older age, family history of prostate cancer, ethnicity (African-American), comorbidity, and symptoms. The diagnosis of prostate cancer is established with histological evaluation of prostate tissue. The systematic biopsy of the prostate under transrectal ultrasound guidance (TRUS) was introduced in 1989, when 6 biopsy cores were taken from the base, middle and apical region of the prostate [22]. Since the procedure was developed, modifications have been instituted to extract a greater number of samples. For example, because prostate cancer is more prevalent in the peripheral zone of the prostate, modern procedures obtain more biopsies from the lateral region of the prostate. Currently, the most common method in Sweden is to sample 10 to 12 biopsy cores from the prostate according to a systematic pattern [23, 24, 25]. Traditionally, an ultrasound device has been used to guide the physician to direct the needle in the areas of the prostate that need to be sampled. This method is widely used today even

though it exhibits low specificity and low sensitivity in the detection of high-grade prostate cancer.

The Gleason Grading System

To evaluate the prognosis of prostate cancer from a prostate biopsy, a microscopic evaluation of the tissue has to be made. The International Society of Urologic Pathology (ISUP) modified Gleason grading system is used to grade prostate biopsies, and reflects the pattern of gland formation and differentiation level of the cells in the prostate (Figure 3.4) [26, 27, 28]. In Figure 3.4 we see that lower Gleason grades are associated by small, tightly packed glands while the cells spread out with higher grades and the glandular architecture loosens [29]. By definition, a total Gleason score is calculated from two diagnostic outcomes (numbers ranging from 3 to 5). The first number is based on the microscopic appearance of the most prevalent graded cell pattern, and the second number is based on the next most prevalent graded cell pattern in the biopsy sample. However, the highest grade needs to be included as either of the two numbers. These numbers are then combined to produce the total Gleason grade and from this, the ISUP grade is produced, resulting in a value between 1 to 5. ISUP grade 1 is equivalent to Gleason score 3+3, ISUP 2 is equivalent to Gleason 3+4, ISUP 3 is equivalent to Gleason 4+3, ISUP 4 is equivalent to Gleason 4+4 and ISUP 5 is equivalent to Gleason 4+5 or higher. Cancers with a higher ISUP grade are more aggressive and have a worse prognosis, indicating higher risk and greater mortality [30, 31, 32].

Gleason score 3+3=6 (Gleason 6), which equals the ISUP grade 1, is the most common prostate cancer diagnosis. According to a cohort study by the National Institutes of Health (NIH), a large proportion (around 50%) of the men diagnosed with prostate cancer in countries with high PSA testing have tumors that have a very low level of metastatic potential (ISUP 1 prostate cancer) [33]. Our research group has observed similar numbers in the Stockholm PSA and Biopsy Registry Database, which includes men diagnosed with prostate cancer in the Stockholm area. Such tumors are unlikely to become metastatic, and mortality rates of men with ISUP 1 prostate cancer are similar to that of men without a prostate cancer diagnosis [34, 35]. Furthermore, in a large American study, Ross *et.al* showed that ISUP 1 prostate cancer is not associated with a metastatic phenotype [36]. Even so, ISUP 1 prostate cancer meets the histological definition of cancer and can affect the patients and cause them considerable psychological stress [37, 38]. Thus, the debate remains whether ISUP 1 prostate cancer should be considered prostate cancer and whether it meets the molecular and genetic criteria for cancer [10, 11, 39, 40].

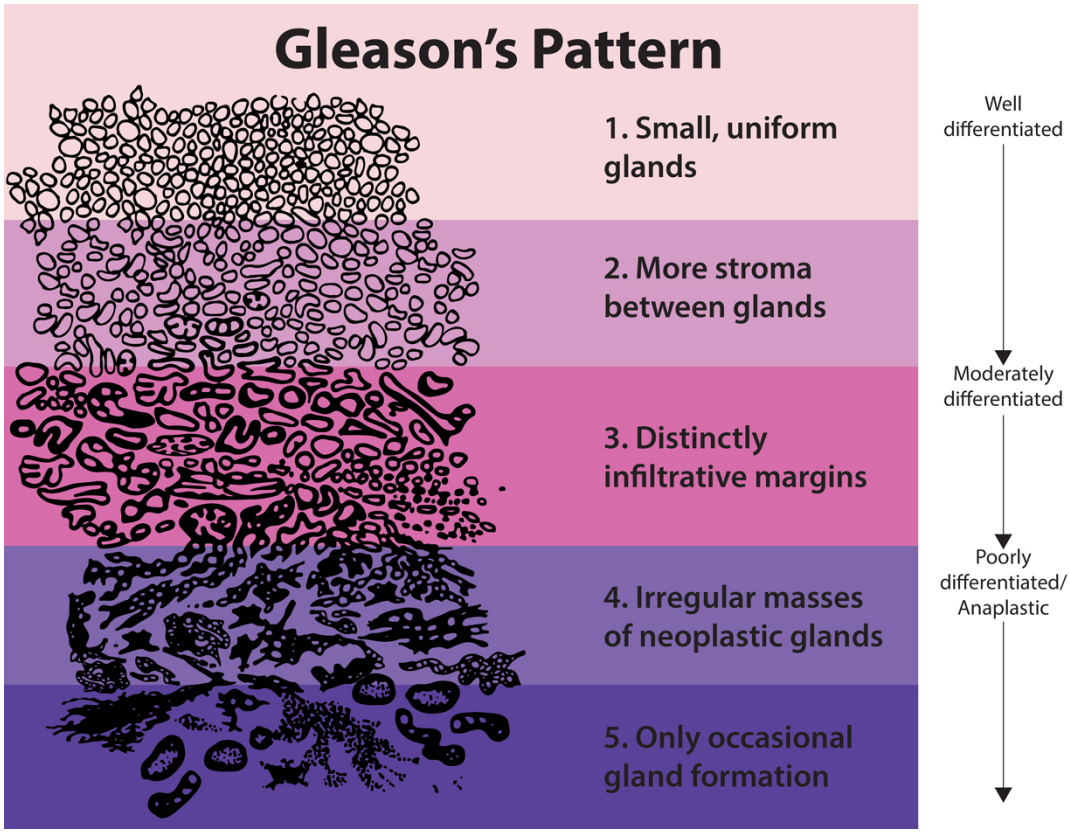


Figure 3.4: Gleason grade. Lower grades are associated with small, closely packed glands. Cells spread out and lose glandular architecture as grade increases [29].

Prostate cancer with ISUP grade 2 or higher is generally a more aggressive type of cancer and has been associated with worse prognosis and higher mortality rates than ISUP 1 prostate cancer [41]. Tumor cells with a higher ISUP grade accumulate in poorly formed glands that are less specificly seperated from the stroma (see Figure 3.4) [29]. The disrupted cell membrane of the tumor cells thus leaks PSA from the prostate into the bloodstream, and therefore the level of PSA is often used to test for prostate cancer. However, it is relevant to note that men with a high-grade prostate cancer normally have high PSA levels while some men with only slightly elevated PSA levels can have very aggressive prostate cancer, but not all men with high PSA levels have prostate cancer.

Prognostic Risk Stratification Systems

There are several prognosis risk stratification systems following prostate cancer diagnosis. According to the Swedish national guidelines, the risk (severity) of the prostate cancer is divided into four categories depending on T-stage, Gleason grade, PSA level and mm cancer in biopsy (see Table 3.1) [42].

Table 3.1: Classification of prostate cancer risk after diagnosis from the Swedish national guidelines [42]

Risk	Definition	Treatment Options
Very low	T1c, Gleason $\leq$ 6, PSA density $<$ 0,15, $<$ 8 mm in $\leq$ 4 out of 8-12 biopsy cores	Appropriate for active surveillance
Low	T1-T2a, Gleason $\leq$ 6, PSA $<$ 10 ng/mL	Appropriate for active surveillance
Intermediate	T2b, Gleason $=$ 7, 10 $\leq$ PSA $\leq$ 19 ng/mL	May benefit from intervention
High	T2c-T3, Gleason $>$ 7 or extensive Gleason $=$ 3+4, PSA $\geq$ 20 ng/mL	Benefit from intervention

Reclassification of ISUP Grade

The reclassification of ISUP grade from systematic biopsy to whole gland pathology on prostatectomy samples is very common and it can have serious consequences for the patient if the diagnosis is erroneous. Reclassification happens for two primary reasons: the biopsy needle can miss the area in the prostate with the highest ISUP grade, and the pathologist grading the whole prostate can disagree with the pathologist grading the biopsy sample (concordance between pathologists on the same biopsy sample is around 30–50% [43]). These classification problems are very common with systematic biopsies, and studies have shown that overall agreement of the ISUP grade when comparing biopsy sample and radical prostatectomy sample from the same individual is only around 35–60% [44, 45, 46].

3.2.3 Initial Treatment

Prostate cancer treatment is a large topic and the focus of an enormous body of research. Here I will only give a very brief and general overview of available treatment options, since it has not been a focus of my research. Currently, there are several options for treatment after prostate cancer diagnosis. Common clinical practice recommendations for men with very low or low risk prostate cancer is active surveillance with no active treatment of the disease. If there is evidence that the cancer is progressing, treatment with curative intent might be recommended [42, 47]. For men with over 10 years expected lifespan and intermediate risk prostate cancer, either radiation or prostatectomy are typically recommended. For men with high risk prostate cancer, either radiation or chemotherapy is commonly recommended, except for men with a metastatic disease, for which androgen suppression treatment is recommended (see Table 3.1) [42, 18].

Active Surveillance

Prostate cancer is a disease with a broad span of prognosis and the typically recommended treatment option for men with ISUP 1 prostate cancer is active surveillance, where the patient is actively followed for signs of disease development rather than intervention. Active surveillance includes repeated PSA testing, MRI and biopsies. If the diagnosis remains static, the men are not actively treated for prostate cancer.

Radical Prostatectomy and Radiation

Men diagnosed with intermediate or high-risk localized prostate cancer without metastases and over 10 years expected lifespan are generally recommended to have a radical prostatectomy or radiation of the prostate [18]. Radical prostatectomy is the surgical removal of the whole prostate gland and the seminal vesicles. The procedure can be performed as an open, laparoscopic or robot-assisted laparoscopic surgery. The surgery has risks and complications following the procedure, with the most common including infections, internal bleeding, erectile dysfunction and urinary incontinence [18]. Radiation of the prostate is also used to treat men with intermediate/high-risk prostate cancer. During this procedure, high-energy radiation beams are aimed at the prostate gland with the goal of destroying the cancerous cells while sparing as much of the normal surrounding tissue as possible [48].

3.3 Improvements in the Prostate Cancer Diagnostic Strategy

The current prostate cancer diagnostic strategy, which includes PSA testing and systematic biopsy analysis for opportunistic screening and diagnosis of prostate cancer, has been shown in many studies to reduce prostate cancer mortality [6, 7, 8]. However, both the PSA test as well as systematic biopsies have demonstrated poor sensitivity and specificity [49, 50], resulting in overdiagnosis, overtreatment, and misdiagnosis of cancer [6, 11, 12, 45]. Therefore, current research focuses on systematically investigating and evaluating a new clinical workflow to screen for and diagnose prostate cancer using risk prediction models, magnetic resonance imaging (MRI) along with targeted biopsy and automated image analysis of biopsy samples (see Figure 3.5). This improved diagnostic pipeline can improve both sensitivity to diagnose high-grade prostate cancer and reduce unintended consequences of screening, e.g. false positive biopsy recommendations, overdiagnosis, and overtreatment.

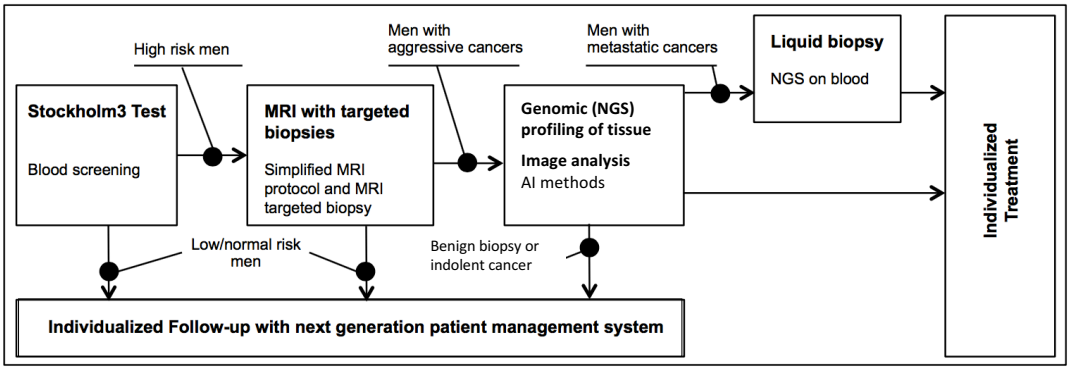


Figure 3.5: One option for improvement in the prostate cancer diagnostic chain starts with a Stockholm3 test, followed by MRI of the prostate and a targeted biopsy. AI assisted technology is then used to grade the biopsy samples with ISUP grade in order to save workload and harmonize ISUP grading. Men with aggressive cancers would then have genomic profiling of the cancerous tissue for individual treatment options. Men with a negative test in the chain will have an individualized follow-up with next generation patient management system.

3.3.1 Prostate Cancer Risk Prediction Models and the Stockholm3 Test

To improve prostate cancer diagnostics, risk prediction models have been developed to systematically recommend men to undergo a prostate biopsy. These models, e.g. Stockholm3, PCPTRC, PBCG, ERSPCRC, 4K, and PHI have shown favourable results compared to the PSA test to detect high-grade prostate cancer [51, 52, 53, 54, 55, 56, 57, 58]. These models have been shown to reduce the number of performed biopsies and decrease detection of ISUP 1 prostate cancer, while being able to detect ISUP ≥ 2 prostate cancer with similar sensitivity.

The STHLM3 study was a prospective, population-based, paired, screen-positive study in men aged 50–69 years without previously diagnosed prostate cancer, which utilizes the Stockholm3 risk prediction model. The Stockholm3 risk prediction model includes a combination of biomarkers (PSA, free PSA, intact PSA, hK2, MSMB, MIC1), genetic polymorphisms (232 SNPs), and clinical variables (age, family history, previous prostate biopsy, prostate exam) to predict the risk of ISUP ≥ 2 prostate cancer. The STHLM3 study consisted of a training (n=11,130) and a validation cohort (n=47,688) of men living in Stockholm County, Sweden. To fit the Stockholm3 model, clinical and genetic data from the training cohort was analysed, which was subsequently evaluated against the validation cohort. All men with a PSA of 1 ng/mL or higher were tested with the Stockholm3 test. The men with a PSA above 3 ng/mL or a Stockholm3 test score above 10% were referred to undergo prostate biopsy. In the study, 7,417 men underwent biopsy, and of those, 1,241 men were diagnosed with ISUP ≥ 2 prostate cancer. The Stockholm3 model performed significantly better in predicting the risk of ISUP ≥ 2 prostate cancer, reducing the number of biopsies by 32% and avoiding 44% of benign

biopsies compared to using a PSA test for referral to biopsy at the same sensitivity for ISUP ≥ 2 prostate cancer. Additionally, the study showed that by using the Stockholm3 test, the number of cancers with ISUP grade 1 could be reduced by 17% by more precisely predicting the risk of high-grade prostate cancer and thus recommending fewer healthy men to undergo biopsy [51]. Thus, the study provided supporting evidence that using a good risk prediction model in addition to PSA testing, can better forecast the probability of ISUP ≥ 2 prostate cancer.

3.3.2 Magnetic Resonance Imaging (MRI) of the Prostate and Targeted Biopsy

Over the last few years, magnetic resonance imaging (MRI) of the prostate has been introduced as part of the clinical assessment of prostate cancer. MRI has the potential to identify areas of the prostate where prostate cancer is likely to reside. Then, using the targeted biopsy, the MRI of the prostate is fused into an ultrasound device, thereby directing the physician toward areas in the prostate where a suspicion of cancer exists. Other available methods are cognitive fusion biopsy or in-bore biopsy. The targeted biopsy method is anticipated to overcome shortcomings of the existing systematic biopsy method, which is performed without precise knowledge of the location of a suspicious lesion. For the reporting of MRI findings, the Prostate Imaging Reporting and Data System (PI-RADS) score has been developed. The score is based on a value from 1 to 5 (given for each lesion), with 1 being most probably benign and 5 being highly suspicious of malignancy [59]. Studies using MRI and targeted biopsies have shown improved diagnostic accuracy of ISUP ≥ 2 prostate cancer, while decreasing both the detection of ISUP 1 prostate cancers and number of performed biopsies [49, 60, 61, 62, 63].

In a new study by Grönberg *et al.*, the Stockholm3-MRI phase 1 study, they combined the Stockholm3 test and MRI targeted biopsies and compared it to current diagnostic methods. The two primary outcomes were the number of performed biopsies and the ISUP-specific number of diagnosed cancers. The results showed that by using a combination of the Stockholm3 test and targeted biopsy, there was a significant 42% reduction in the number of performed biopsies and 46% reduction in the number of diagnosed ISUP 1 prostate cancers, while detecting the same number of ISUP $\geq$ 2 prostate cancers. Thus markedly improving the diagnostic specificity compared to current practice [62].

3.3.3 Automated Image Analysis of Prostate Biopsy Samples

The possible next step in improving the prostate cancer diagnostic pipeline is to use artificial intelligence for the improvement of image analysis of ISUP grading of prostate biopsy samples. Currently there is a severe shortage of uro-pathologists well trained in

grading biopsy samples, which might result in the earlier mentioned variability between pathologists on the same biopsy sample (reclassification of ISUP grade). With a more objective evaluation of biopsy samples, using artificial intelligence assisted prostate cancer pathology these negative effects can be reduced. The aim is to reduce the variability in pathology assessment, increase throughput of analysis (shorten waiting time for patient diagnosis) and to improve prediction of the correct disease status of prostate cancer compared to current pathology procedures.

Chapter 4

Data Material

In this thesis, I have used three data sets: the observational Stockholm PSA and Biopsy Registry, data from the prospective and population based STHLM3 diagnostic trial, and data from the Stockholm3-MRI phase 1 study.

4.1 The Stockholm PSA and Biopsy Registry

In Study I we used data from the Stockholm PSA and Prostate Biopsy Registry, which links PSA, biopsy and prostate cancer diagnosis data in Stockholm, Sweden. Information on PSA tests was collected from the three clinical chemistry laboratories that performed all PSA tests in Stockholm, from 2003 to 2015. Data on prostate biopsies were collected from three pathology departments in Stockholm and family history of prostate cancer was obtained from the Swedish Multi-generation Register. By linking to the Regional Prostate Cancer Register and the Swedish National Cancer Register we obtained data on tumor stage, Gleason score, and mode of detection (PSA detected or symptomatic detection). On November 23, 2015, the Stockholm PSA and Biopsy Registry included data from 448,000 men and 1.8 million PSA tests. The study was approved by the local ethics review board. One of the strengths of my thesis is having access to this data set since it is a unique registry describing the PSA testing and biopsies in a whole region with around 2 million inhabitants. Like all registry data, the data set has its limitations, with data that has to be cleaned and may contain some registration errors.

4.2 The STHLM3 Cohort

The STHLM3 study was a prospective and population-based diagnostic study including 59,149 men aged 50–69 years without prostate cancer, randomly invited from the Swedish Population Register. In total 7,417 men were biopsied. The STHLM3 dataset is unique, including information of a combination of plasma protein biomarkers, genetic markers, and clinical variables on all participants including detailed biopsy information

as well as a prostate exam (digital rectal exam and prostate volume) on the men who were biopsied. We used data from the STHLM3 cohort in Study II and IV.

Currently, 780 men in the STHLM3 cohort have had a radical prostatectomy as a treatment for high-risk prostate cancer. We have collected extensive pathology information on those men and compiled the STHLM3 Radical Prostatectomy Cohort. The data was used in Study IV of my thesis.

The limitations of the study are that so far there is limited follow-up time, and most of the men in the study were of Northern European descent. Currently there are ongoing validation studies to broaden the scope and generalize the use of the Stockholm3 test.

4.3 The Stockholm3-MRI Phase 1 Cohort

In Study III we used data from the Stockholm3-MRI Phase 1 study, which was a prospective, multi-center, paired diagnostic study [64]. Patients were recruited between 1st of May 2016 to 1st of June 2017 from three sites: Stockholm, Sweden and Oslo and Tønsberg, Norway. Men aged 45–75 years without a previous prostate cancer diagnosis and referred to the sites for a prostate cancer diagnostic workup (prostate biopsy or pre-biopsy MRI) were eligible for inclusion (see Figure 4.1).

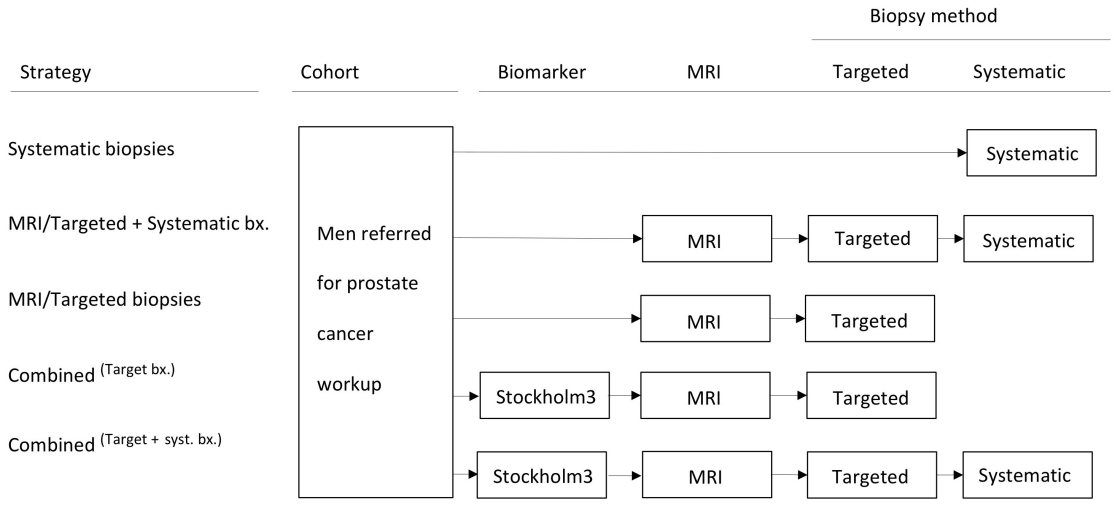


Figure 4.1: Diagnostic strategies for the detection of prostate cancer in the Stockholm3-MRI phase 1 study [62].

In total, 532 men participated in the study and they all underwent blood sampling, MRI and a combined systematic/MRI targeted biopsy procedure depending on the MRI result. The main goal of the study was to compare the performance of MRI-TRUS fusion targeted biopsy and/or systematic biopsy with and without the requirement of a prior positive Stockholm3 test. This study investigated an important link in the future prostate cancer diagnostic chain, and one of its strengths is combining the information from the Stockholm3 test with the results from MRI and targeted biopsies.

Chapter 5

Methods - Prediction Models in Medicine

Prediction models are becoming increasingly important in our society and personal lives. Their purpose can range from predicting the risk of a disease in health care to predicting the risk of a person recommitting a crime in the judicial system. These models are most often used in the background without our knowledge and they can be very important for the development of great improvements as well as harmful and discriminating if used unfairly.

Within the field of medicine, the use of prediction models is increasing. They can be effective in predicting the risk of a disease in order to decide if additional testing is needed. If the risk of a disease is low, additional invasive tests can be avoided but when the risk is high, further diagnostic workup can be necessary for the patient. The ideal would be to have a disease status test that predicts with perfect accuracy the true disease status of the patient. However, no such tests are available and precise tests are often invasive and can cause side effects that should be avoided if possible (such as biopsy or surgery to determine cancerous tissue). Thus less invasive risk prediction models are needed to aid physicians in their decision of an invasive intervention to determine disease status.

Examples of the use of prediction models in medicine are early detection of a disease (a blood based, urine based or radiology based screening test), automating the process of image analysis for medical purposes and individualized treatment based on a person's genetics, environment and lifestyle. The current era of evidence based medicine requires the development of more advanced techniques and methods to individualize the whole diagnostic chain based on each person's attributes, from *screening* to *diagnosis* and *treatment*.

5.1 Model Development

5.1.1 Statistical Models for Prediction

The first step of the risk prediction model development is the model selection. The model we select to represent our data in our risk prediction model, has to be based on the type of data we have and the type of outcome we want to predict the risk for. The outcome we want to predict can be continuous, binary, categorical, or ordered categorical and we will have to choose different models for each type of outcome.

For prediction of **continuous variables**, the most commonly used model is the linear regression model where the parameters of the model are selected by the principle of least squares; minimizing the sum of the squares of the differences between the observed dependent variable and those predicted by the model.

To predict the risk of **binary variables** in medical outcomes, the logistic regression model is very useful, especially for diseased and non-diseased outcomes. It models a binary outcome variable using the logistic function and is used to test for potential associations between a set of given variables and a binary or categorical outcome, as well as predicting the risk of the outcome variable based on one or more predictors. In a logistic regression, we use the logistic link function to restrict the predictions between 0 and 1. The model is the linear function of the logistic transformation of the probability p of the outcome y or $P(y = 1)$ using the logistic function:

$$\text{logit}(p) = \log(odds(p)) = \log(\frac{p}{1-p}) = \alpha + \beta_i x_i \tag{5.1}$$

where α is the intercept, β_i are the estimated regression coefficients and x_i are the predictors. The most common way to estimate the coefficients is using the maximum likelihood method, but other less used methods are available. The interpretation of the β coefficients is the same as with other regression models. It indicates the difference in the log-odds for a one unit increase in its predictor holding all other predictors in the model constant. The odds ratio is calculated as the exponent of the β_i coefficients. The predicted probability p is calculated by:

$$p(y = 1) = \frac{e^{(\alpha+\beta_i x_i)}}{1 + e^{(\alpha+\beta_i x_i)}} = \frac{1}{1 + e^{-(\alpha+\beta_i x_i)}} \tag{5.2}$$

The maximum likelihood method can be used to optimize the logistic model, i.e. to

find the most likely value of the model's parameters given the data. The log-likelihood is used because it is computationally easier to maximize than the likelihood. It is calculated as the sum over all subjects of the distance between the natural log of the predicted probability p for the binary outcome to the actually observed outcome y [65]:

$$LL = \sum_{j=1}^{N} y_j \cdot \log(p_j) + (1 - y_j) \cdot \log(1 - p_j) \tag{5.3}$$

where y is the binary outcome variable and p is the predicted probability for each subject for the outcome.

From equation 5.1 we can calculate the odds ratio (OR) comparing the odds of $y = 1$ when $x = 1$ to when $x = 0$:

$$OR = \frac{odds(x+1)}{odds(x)} = \frac{\exp(\alpha + \beta(x+1))}{\exp(\alpha + \beta x)} = \exp(\beta) \tag{5.4}$$

There are multiple other models available to derive predictions for binary outcome variables. These are for example neural networks, regression tree and support vector machine.

In Study I and II a multinomial, multivariable logistic regression was used to predict the risk of three different outcomes of biopsy: benign, ISUP 1 prostate cancer and ISUP ≥ 2 prostate cancer. The outcome variable is a **categorical variable** and takes three different values. For predictor x, each response level follows a logistic regression model for x, specifying a selected level as the reference. Thus for level j of the outcome compared to reference level s of the outcome, the model is:

$$log\left(\frac{P(y=j)}{P(y=s)}\right) = \alpha + \beta_i x_i \tag{5.5}$$

Thus, β_i is the log odds ratio comparing the odds of outcome $y = j$ to the odds of $y = s$ for a unit increase in x_i [66].

Ordinal outcomes are also quite common in medical studies and to predict those types of outcomes we can use the proportional odds logistic regression model. In this

extension of the logistic regression model, we assume a common set of regression coefficients across all levels of the outcome and intercepts are estimated for each level.

5.1.2 Overfitting and Underfitting in Prediction Models

When developing a good prediction model, a primary interest should be the model's performance in prediction of data outside the sample study. One of the issues in the development of prediction models is called **overfitting**, i.e. the predictions fit the data well in our sample study, but the model does not predict well on new data outside the study sample. Overfitting causes optimism in the evaluation of our model's performance and it can bee seen as the *apparent* performance on our data sample compared to the *true* performance of the model when evaluated on a population sample. Optimism of a prediction model is defined as the difference between the apparent performance and the true performance (see equation 5.6) [65].

$$Optimism = Apparent\ performance - True\ performance \tag{5.6}$$

When comparing prediction models, the model's optimism is one measure we can use for comparison. Models with higher optimism are possibly overfitting the data and can be improved.

When the complexity of our models is not capturing important distinctions and patterns in our data that results in a high prediction error in our training and validation data as well as external data it is called **underfitting**. To solve the problem of underfitting, adding more features to our model is effective. Thus, to find the optimal model, we need to keep midlevel complexity in the models to minimize both underfitting and overfitting (see Figure 5.1). Methods such as bootstrapping and cross-validation are available to help us quantify and decrease overfitting in our models.

5.1.3 Ensemble Modeling

Ensemble modeling is a statistical learning method that uses several different models to improve the stability and predictive performance of our prediction model, either by using different types of models or different training data sets. Below I will briefly describe two ensemble modeling methods, stacking and boosting.

Stacking

For the purpose of improving the Stockholm3 predictions in Study IV, using two different logistic models, one model trained on the biopsy data and one model trained

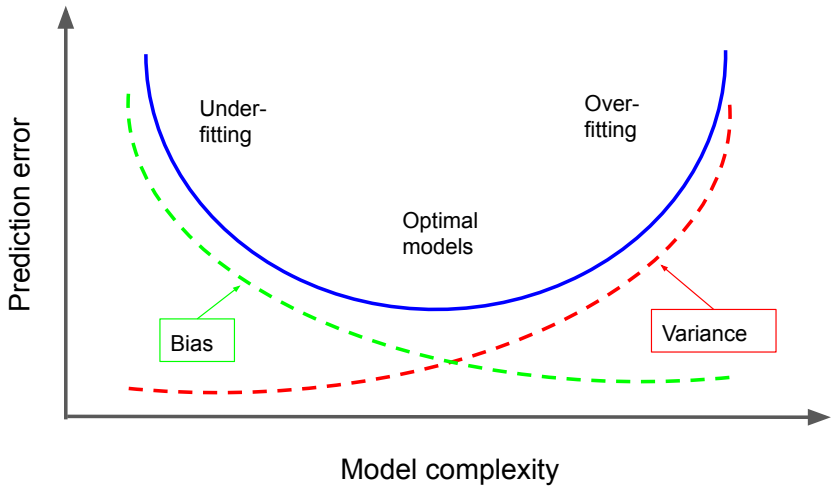


Figure 5.1: Optimal models minimize underfitting and overfitting. Overfitting increases with model complexity while underfitting decreases.

on the prostatectomy data, we explored a method called *stacking*, which combines many different models to improve the predictions. For a set of models $m = 1, 2, ..., M$ and given predictions $\hat{f}_1(x), \hat{f}_2(x), ..., \hat{f}_M(x)$ for those models, $\hat{f}_m^{-i}(x)$ is the prediction at x, using model m of M models, with the ith training observation removed. We obtain the stacking estimate of the weights from the logistic regression of y_i on $\hat{f}_m^{-i}(x_i), m = 1, 2, 3, ..., M$. Stacking weights are then given by:

$$\hat{w}^{st} = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{N} \left[y_i - \sum_{m=1}^{M} w_m \hat{f}_m^{-i}(x_i) \right]^2 \tag{5.7}$$

which gives us the final predictions: $\sum_m \hat{w}_m^{st} \hat{f}_m(x)$ [67]. The method did not increase the performance of our prediction model in study IV and thus the results of the analysis are not presented in the paper.

Boosting

The main principle of boosting algorithms for prediction models is to fit a sequence of models to weighted versions of the data. A boosted classifier is in the form:

$$F_T(x) = \sum_{t=1}^{T} f_t(x) \tag{5.8}$$

where each $f_t(x)$ is a weak learner that takes x as input and returns a value that is the predicted class of the specific weak learner. One example of a boosting method is the

AdaBoost algorithm introduced by Freund and Schapire [68] in 1996. In the AdaBoost algorithm, the examples that were misclassified in the subsequent classifiers are given more weight so that the weak learner is forced to focus on the hard examples in the training set [69]. Then the predictions are combined with a weighted classification to produce a final prediction. The boosting method did not improve our predictions in the Stockholm3 model and thus the results of the analysis were not presented in Study IV.

5.2 Internal Model Validation

When choosing the best model to use for our data we need to have two goals in mind; first, the model selection and second, the model assessment: When we have chosen a model and its parameters, we need to estimate its prediction error on new data. If we have a large data set to address both issues it is optimal to divide the data sets into three: a training set, a validation set, and a test set. The training set is used to train the model; the validation set is used to estimate the prediction error for model selection and finally we use the test set to assess the generalization error of the chosen model. Optimally a test set should be set aside before the analysis begins and then only used once at the end of the data analysis for validating the selected model. The methods available for model validation are e.g. apparent validation, split-sample validation, cross-validation and bootstrapping. As cross-validation and bootstrapping are the most widely used I will describe them in more detail.

K-fold Cross-validation

In a data sparse situation, which is often the case with clinical data, it is not optimal to remove big parts of the data for validation and testing of the model. Then K-fold cross-validation, an internal validation technique, is very useful to estimate the expected prediction error. It can be used to evaluate how well the risk prediction model will work in practice on an external data set using the same data set used to train the model as well as to decrease overfitting of the data.

To handle the scarce data problem, in K-fold cross-validation we use a part of the available data to fit the model and another part to test it [67]. The data is split into K different data sets of similar size (see Figure 5.2 where $K = 5$). Then the model is trained on the other $K - 1$ data sets (the blue boxes) and the prediction error is estimated on the kth part (the red box). This is repeated for $k = 1, 2, 3, ..., K$ and to produce the cross-validation estimate of the prediction error, the K error estimates are accumulated over all K folds.

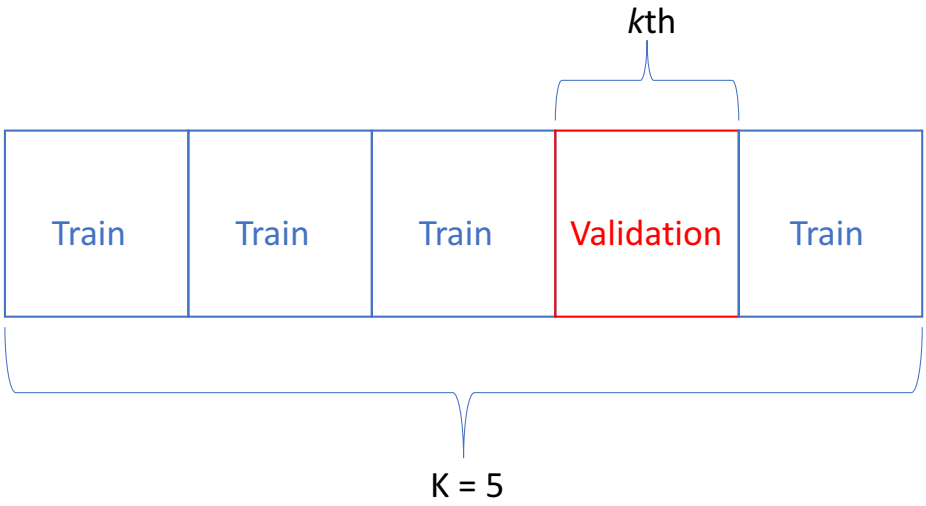


Figure 5.2: K-fold cross-validation: Total data set is divided into K parts of equal size, $K-1$ parts are used to train the model which is then tested on part k of the data to estimate prediction error in our model.

Bootstrapping

Bootstrapping is a method that relies on random sampling with replacement. The word bootstrapping comes from the phrase to *pull oneself up by one's bootstrap* based on a German tale of the notorious Baron Münchhausen who was famous for the exaggerated tales of his own adventures. According to the tale, he was able to pull himself out of a swamp by his own hair. In a later version of this story, he used the bootstraps of his shoes to pull himself out of the sea [70]. However, in statistics, the Baron does not pull himself up by his bootstraps but the word is used for a method to estimate the sampling distribution of an estimator by resampling with replacement [71](similar to the Baron's achievements but maybe not as unbelievable as his hair pulling!). Contrary to the tale, the method is widely applicable and can provide valuable insight for statistical inference to deduce the probability distribution of a population.

The method works as follows: With a total sample size of n, we draw with replacement a sample size m from n (typically $m = n$) called the bootstrap sample. We repeat this procedure many times (often a few thousand) and then we calculate a statistic that we are interested in for each bootstrap sample to estimate the variance. The bootstrap theory can be used to make computerized calculations of basic statistical concepts, like e.g. confidence intervals [71].

Bootstrapping can also be used to quantify the optimism (Equation 5.6) of a prediction model [72] as well as to calculate the optimism-corrected performance [65].

With a bootstrap variant, a model is repeatedly fitted in bootstrap samples, and then the performance is evaluated in the original sample. To estimate the optimism, we use the bootstrap:

$$Optimism = Bootstrap\ performance - Test\ performance \tag{5.9}$$

Then we can calculate the Optimism–corrected performance:

$$Optimism{-}corrected\ performance = Apparent\ performance\ in\ sample - Optimism \tag{5.10}$$

5.3 Model Evaluation

The predictive abilities of clinical risk prediction models are important to evaluate. The most common and recommended statistical tools to evaluate these predictive abilities evaluate discrimination, calibration and clinical relevance of the models.

5.3.1 Discrimination

Discrimination evaluates how well the predicted risk from the model distinguishes between patients with and without disease. The most commonly used measure for discrimination is the concordance (c)-statistic and for a binary outcome model the c-statistic is the area under the receiver operating characteristic (ROC) curve (AUC). It is based on the specificity and sensitivity of our model to predict the risk of disease and plots of the sensitivity vs. specificity for different cutoff thresholds for the probability of the disease predicted by our risk prediction model.

Sensitivity is the number of true positive predictions divided by the number of those with the disease (see Equation 5.11) and the specificity is the number of true negative predictions divided by the people without the disease (see Equation 5.12). To classify a patient as having the disease, we have to choose a threshold and if the predicted risk exceeds that threshold, the patient is classified as having the disease. If the predicted risk is below that threshold, the patient is classified as not having the disease. The ROC curve is a plot of the sensitivity vs. specificity over the whole range of thresholds from 0% to 100% (see an example of a ROC curve of a prediction model in Figure 5.3 showing the sensitivity and specificity for 4 different cutoff thresholds).

$$Sensitivity = TP/n_{disease} \tag{5.11}$$

$$Specificity = TN/n_{notdiseased} \tag{5.12}$$

A risk prediction model with a ROC curve that has an AUC close to 1 is a very good prediction model, meaning that the model almost perfectly discriminates between individuals with and without the disease, while a model with an AUC close to 0.5 is no better than randomly assigning the individuals into healthy and diseased groups. We can interpret the AUC as the probability that a random patient with the outcome is given higher probability of the outcome by the model than a randomly chosen patient without the outcome [65].

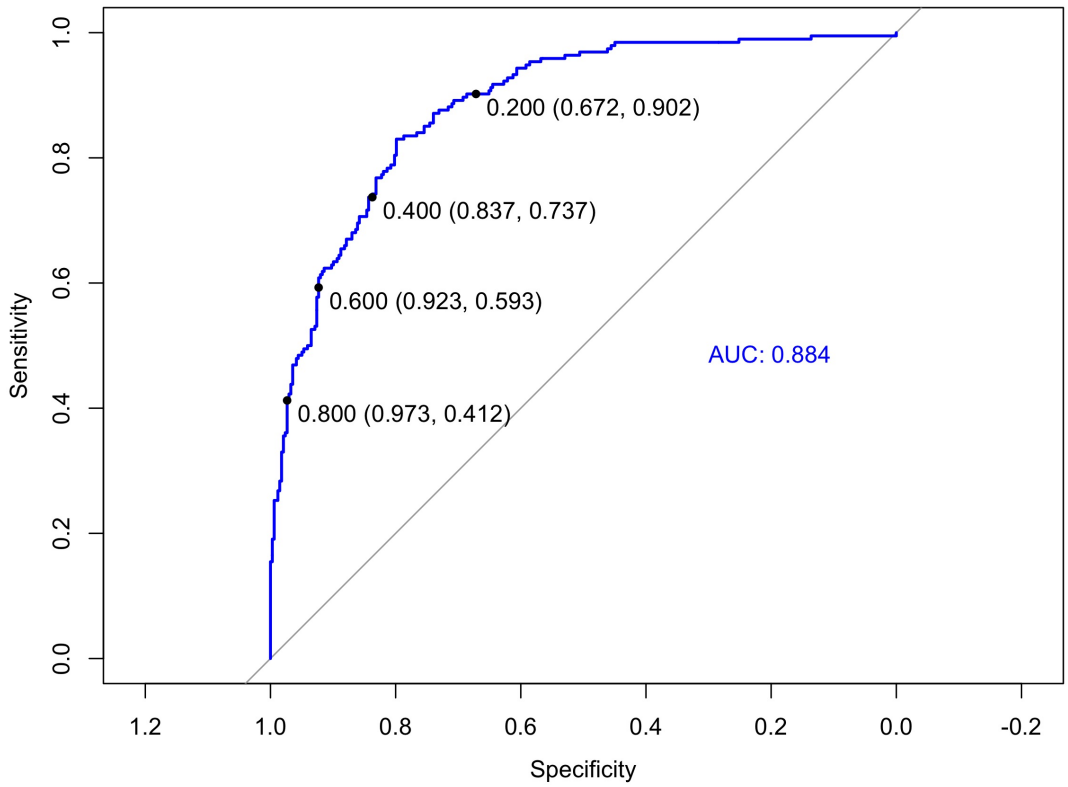


Figure 5.3: Example of a ROC curve of a clinical risk prediction model, the S3M-MRI model developed in Study III. This ROC curve has an AUC of 0.884 and points on the ROC curve represent the specificity and sensitivity for 4 different cutoff thresholds in our model.

5.3.2 Calibration

Calibration Plot

The calibration of prediction models evaluates the agreement between the predicted risks from the model and observed values in the data. Thus, if we predict that a patient has a 10% risk of prostate cancer, the observed frequency of prostate cancer should be 10 out of 100 patients with the same covariate values of the one on which the risk

was predicted [65]. Calibration assessments include **calibration plots**, where we plot the predicted values on the x-axis and observed outcomes on the y-axis. For regression models, this is a scatter plot, while for logistic regression models when the outcomes are 0 and 1 this is not possible. To estimate the observed probabilities relative to the predicted probabilities we can us a smoothing technique (such as loess, kernels or bin smoothing).

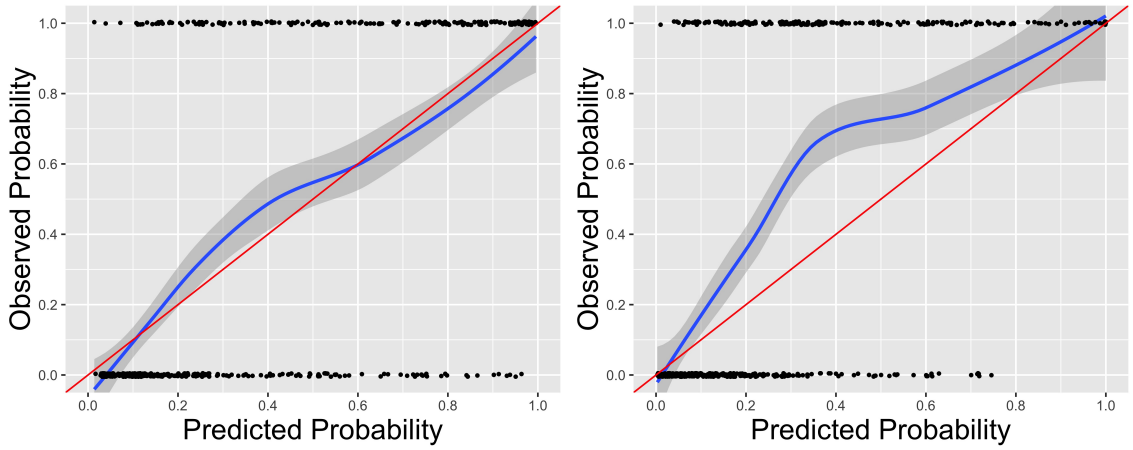


Figure 5.4: Examples of calibration plots for two clinical risk prediction models, with predicted risks on the x-axis and observed outcomes on the y-axis.

Figure 5.4 shows examples of calibration plots for two prediction models. The plots show both the observed outcomes (black dots) and the smoothed estimated outcomes (the blue line) for the observed probabilities of a clinical risk prediction model with a binary outcome. The observed outcomes lie at y = 0 and y = 1. When a model is perfectly calibrated (observed probabilities match exactly the predicted probabilities) then the blue line follows the 45° red line. In Figure 5.4, the model to the left is better calibrated than the model to the right. When we plot the calibration for a model on the data that we used to develop the model, that is the *apparent* calibration, because in model development, the average of the outcomes is the average of the predictions ($Mean(Y) = Mean(\hat{Y})$) [65]. Applying the model to external data can result in worse calibration.

Calibration-in-the-large

Calibration-in-the-large and Calibration slope are two calibration measures that can be used to estimate how well a prediction model is calibrated on external data and to compare the model to other existing prediction models. Calibration-in-the-large is the difference between the mean of the predictions ($\hat{Y}$) and the mean of the observations (Y_{new}):

$$Calibration-in-the-large = Mean(Y_{new}) - Mean(\hat{Y}) \tag{5.13}$$

This applies for calibration of linear regression models and we can test the difference in the means with a one-sample t-test [65]. For calibration of logistic regression models we can make a comparison with an odds ratio:

$$OR = \frac{odds(Mean(\hat{Y}))}{odds(Mean(Y_{new}))} = \frac{\frac{Mean(\hat{Y})}{1-Mean(\hat{Y})}}{\frac{Mean(Y_{new})}{1-Mean(Y_{new})}} \tag{5.14}$$

To test the statistical difference in a logistic regression we need to compare $logit(Y_{new} = 1)$ to $logit(\hat{Y} = 1)$. We can write that:

$$\begin{aligned} logit(Y_{new} = 1) - logit(\hat{Y}) &= a \\ logit(Y_{new} = 1) = a + logit(\hat{Y}) &= a + offset(linearpredictor) \end{aligned} \tag{5.15}$$

We can test statistical significance of the intercept being equal to zero vs. being not equal to zero with a Wald test or likelihood ratio (LR) test.

Then we estimate the calibration slope from the recalibration model:

$$logit(Y_{new} = 1) = a + b_{overall} * logit(\hat{Y}) = a + b_{overall} * linearpredictor \tag{5.16}$$

The miscalibration of the model (the deviation of the slope from 1) can be tested by:

$$logit(Y_{new} = 1) = a + b_{miscalibration} * linearpredictor + offset(linearpredictor) \tag{5.17}$$

The $b_{miscalibration}$ is the slope coefficient and its deviation from 1 can be tested with a Wald test or LR test.

Goodness-of-fit Tests

To test the calibration of a prediction model with binary outcomes, the **Hosmer-Lemeshow (H-L) test** has been used [73]. Goodness-of-fit of a prediction model is the ability of the model to fit a given set of data. Specifically the H-L test calculates if the predicted probabilities match the observed probabilities in population subgroups of our data. The test has shown to have several limitations, e.g. it only tests for overall calibration error and not for any particular lack of fit like quadratic effects and it does not properly take overfitting into account [74]. Therefore, the test is usually not recommended and I have not used it in my studies to test for calibration.

5.3.3 Decision Curve Analysis

Discrimination and calibration are very important features for evaluating the predictive performance of clinical prediction models. However, when comparing two models, one can have better discrimination and the other better calibration. So how do we decide which model is a better clinical prediction model? Decision curve analysis was developed to overcome these limitations of available evaluation methods. *Vickers et al.* introduced the method to calculate the net benefit (NB) of a prediction model as the key part of the decision curve analysis (DCA) [75].

Net benefit of a prediction model is defined as:

$$NB = \frac{TruePositives}{n} - \frac{FalsePositives}{n} * \frac{p_t}{1-p_t} \tag{5.18}$$

where p_t is the probability threshold for declaring a patient positive on the test, $True-Positives$ is the number of patients with the disease at threshold p_t, $FalsePositives$ is the number of patients predicted positive at threshold p_t that do not have the disease, and n is the total number of patients. This number can be thought of as net profit, or income minus expenditure [76].

It is important to evaluate the net benefit over a range of thresholds p_t since there is no one correct threshold for prediction models. Further, to interpret the NB we need to compare the NB of a prediction model with two default strategies, where we "*treat all*" or "*treat none*". *Treat none* always has NB equal to zero and *treat all* is evaluated at reasonable values of p_t. For p_t above prevalence, *treat all* has a lower NB than *treat none* and for p_t below prevalence the *treat all* has a higher NB than *treat none*. For a model to be clinically beneficial it has to have higher NB than *treat all* and *treat none* [76].

Figure 5.5 is an example of a decision curve analysis where NB is plotted for two prediction models against different thresholds p_t from 0 to 0.5. From this analysis we see that model 2 has a NB above the default strategy to treat all if the risk threshold is higher than 0.05 while model 1 has a lower NB than treating all from that threshold.

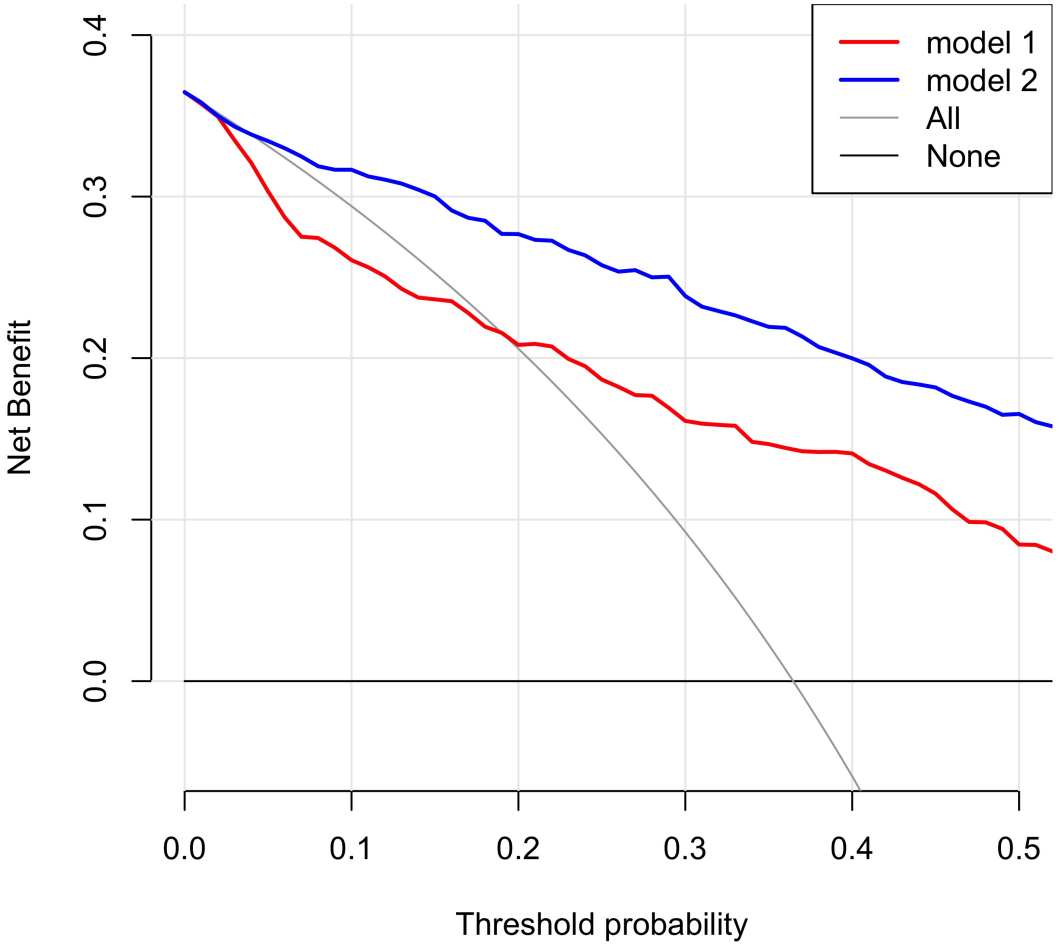


Figure 5.5: Decision curve analysis net benefit curve for two clinical risk prediction models, with net benefit plotted on the y-axis for different values of the threshold p_t on the x-axis.

When a model has a lower NB than any default strategy, the model can be considered clinically harmful. Well calibrated models can not be harmful while miscalibrated models can be harmful. An example of that would be a prostate cancer risk model (such as model 1 in Figure 5.5) which is miscalibrated and predicts the risk of high-grade prostate cancer too low. Using such a risk prediction model in a clinical setting would

result in not sending men to biopsy that likely have the disease, and thus it is harmful for the patient. A model with good discrimination and poor calibration, can be improved by re-calibrating the model to fit external data better. That would result in a model with higher net benefit then before.

5.3.4 External Validation Studies

When developing clinical prediction models, the model is trained, validated and evaluated on patients that are a small sub sample of the total population of patients. The data we train our models on, most often originates from only one country, with most patients of the same race and possibly members of a similar gene pool. Then the big question is: How generalizable is our model in a completely different sub sample of patients? To be sure that our risk prediction model is good, we need to validate the model on external data to test the model's performance on different patients. A model can do extremely well at predicting the risk of a disease for the same type of patients in the data the model was trained on, but when exposed to different patients, the model possibly does not recognize these new predictors and how they are related to the outcome.

Chapter 6

Results

In summary the main findings of my four studies were:

Study I We observed that men with PSA above 1 ng/mL, have an increased risk of being diagnosed with ISUP ≥ 2 prostate cancer with longer than annual PSA testing intervals. However, we also showed that annual PSA testing intervals increase the cumulative probability of having a negative biopsy compared to biennial and triennial testing intervals.

Study II We showed that the risk of ISUP 1 prostate cancer is not significantly associated with PSA and age at time of diagnosis.

Study III A unified S3M-MRI risk prediction model, using the Stockholm3 score and MRI of the prostate to predict the risk of ISUP ≥ 2 prostate cancer is superior to the Stockholm3 model and MRI alone. However, the improvement was small compared to the sequential use of first Stockholm3 and then MRI, which results in fewer MRI examinations and is simpler from a clinical workflow perspective.

Study IV Reclassifiction of ISUP grade between biopsy and radical prostatectomy specimens has little effect on the true predictive performance of risk prediction models, but can lead to a large decline in their apparent prective performance.

6.1 Study I

In Study I we analyzed the benefits and harms of different lengths of PSA testing intervals for men aged 50–74 years. The benefits were decreased risk of ISUP ≥ 2 prostate cancer in biopsy and the harms were increased risk of a false-positive biopsy. We calculated the risk ratio (RR) of ISUP ≥ 2 and ISUP 1 prostate cancer vs. a benign outcome at prostate biopsy and the 12-year cumulative probability of having a negative biopsy by PSA testing intervals, PSA level, age and family history of prostate cancer.

The main analysis showed that men with PSA above 1 ng/mL had increased risk of

being diagnosed with ISUP ≥ 2 prostate cancer when tested with PSA testing intervals over 1 year and the RRs ranged from 1.4 to 3.2 depending on testing intervals and PSA level (see Table 6.1). Our results also showed that men with PSA below 1 ng/mL were at low risk of being diagnosed with ISUP ≥ 2 prostate cancer irrespective of PSA testing intervals, only 5% of the men in our study with PSA below 1 were diagnosed with ISUP ≥ 2 prostate cancer. Age and family history status did not affect the results.

Table 6.1: Risk ratios for ISUP ≥ 2 prostate cancer compared with benign biopsy by pre-index PSA value and PSA testing intervals using 1-year testing interval as baseline

Outcome	PSA value	2 vs 1 yr	3 vs 1 yr	4 vs 1 yr	5-8 vs 1 yr
ISUP ≥ 2	0-1	0.6(0.2 to 1.9)	0.7(0.2 to 3.3)	0.9(0.2 to 4.1)	2.8(1.3 to 6.3)
	1-3	1.3(1.1 to 1.7)	1.7(1.3 to 2.2)	1.8(0.2 to 2.4)	2.5(2 to 3.1)
	3-5	1.3(1.1 to 1.7)	1.7(1.3 to 2.2)	1.8(0.2 to 2.4)	2.5(2 to 3.1)
	5-10	1.4(1.2 to 1.7)	1.2(0.9 to 1.5)	1.7(1.3 to 2.2)	1.6(1.3 to 2.1)

We also calculated the cumulative probability for men without prostate cancer aged 50 and 60 years of having a negative biopsy during 12 years of PSA testing (see Figure 6.1). We found that with shorter PSA testing intervals, the cumulative probability of a negative biopsy was twofold when tested annually compared to biennially and threefold when tested annually compared to triennially. In summary, the cumulative probability of receiving at least one negative biopsy after 12 years of PSA testing 1) decreased with longer testing intervals, 2) increased with pre-index PSA level and 3) was only slightly affected by age and family history of prostate cancer.

6.2 Study II

In Study II, we studied the association of PSA level and age with ISUP 1 and ISUP $\geq$ 2 prostate cancer, respectively. We included 72.996 biopsy cores from 6.083 biopsied men aged 50–69 years in the STHLM3 study. In the overall ISUP grade, the lower ISUP grade can be masked by a higher ISUP grade, therefore we studied the associations for both overall ISUP grade and ISUP grade on each biopsy core. Our results showed that the risk of ISUP 1 prostate cancer was not significantly associated with PSA level in biopsy, neither on an overall ISUP grade level nor on a biopsy core level (see predicted risks in Figure 6.2).

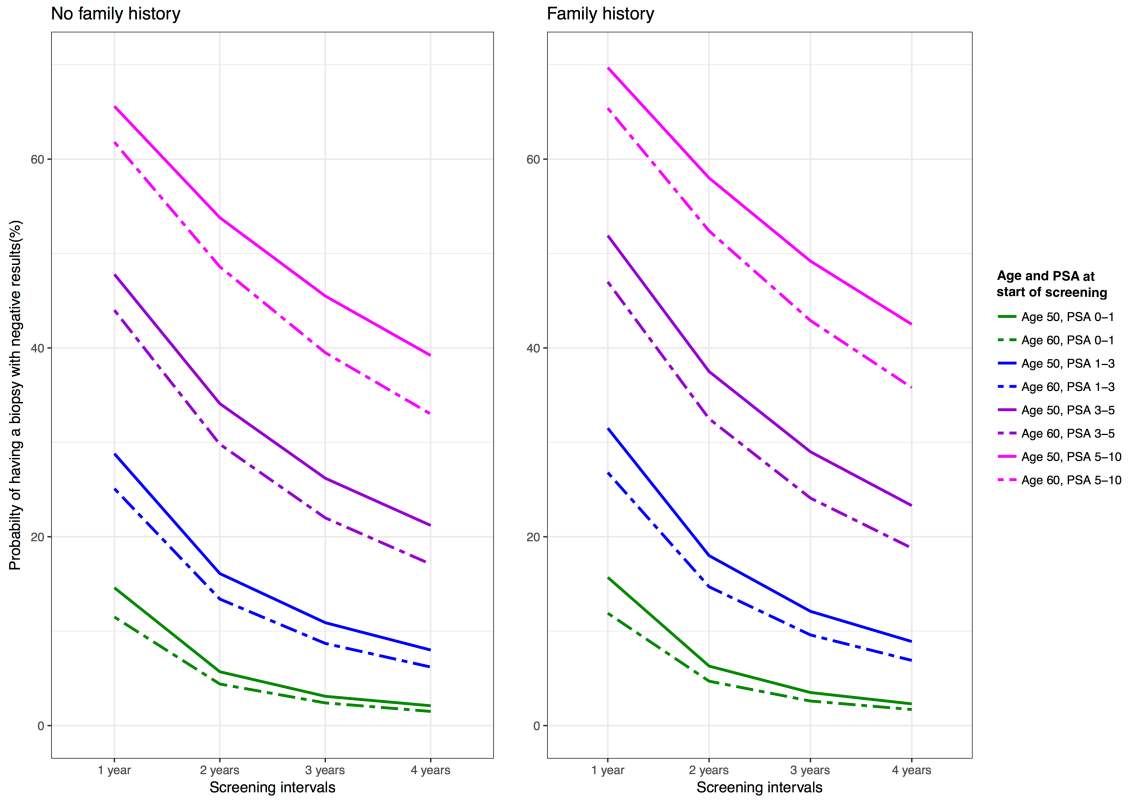


Figure 6.1: Cumulative probability of a negative biopsy for different PSA testing intervals by age and PSA level at start of testing and family history status.

Contrary to ISUP 1 prostate cancer, our results showed that the risk of ISUP ≥ 2 prostate cancer was significantly associated with the PSA level. The analysis showed similar results for the association of the risk of ISUP 1 and ISUP ≥ 2 prostate cancer and age. There was no statistically significant association between age and ISUP 1 prostate cancer in contrast to the strong and significant association between age and ISUP ≥ 2 prostate cancer.

6.3 Study III

In Study III we developed a unified prostate cancer risk prediction model (S3M-MRI) that combined the Stockholm3 score (using biomarkers, clinical variables and a genetic score) and the PI-RADS score from MRI of the prostate. To develop and test our model, we used data from the Stockholm3-MRI phase 1 study, including 532 men without prostate cancer that were referred to a urologist at three sites in Stockholm, Oslo and Tönsberg between 2016 and 2017. We then compared the predictive abilities of the S3M-MRI to the Stockholm3 test and PI-RADs score separately with respect to model discrimination, calibration and net benefit. We then compared clinical outcomes with five different diagnostic strategies using the three models.

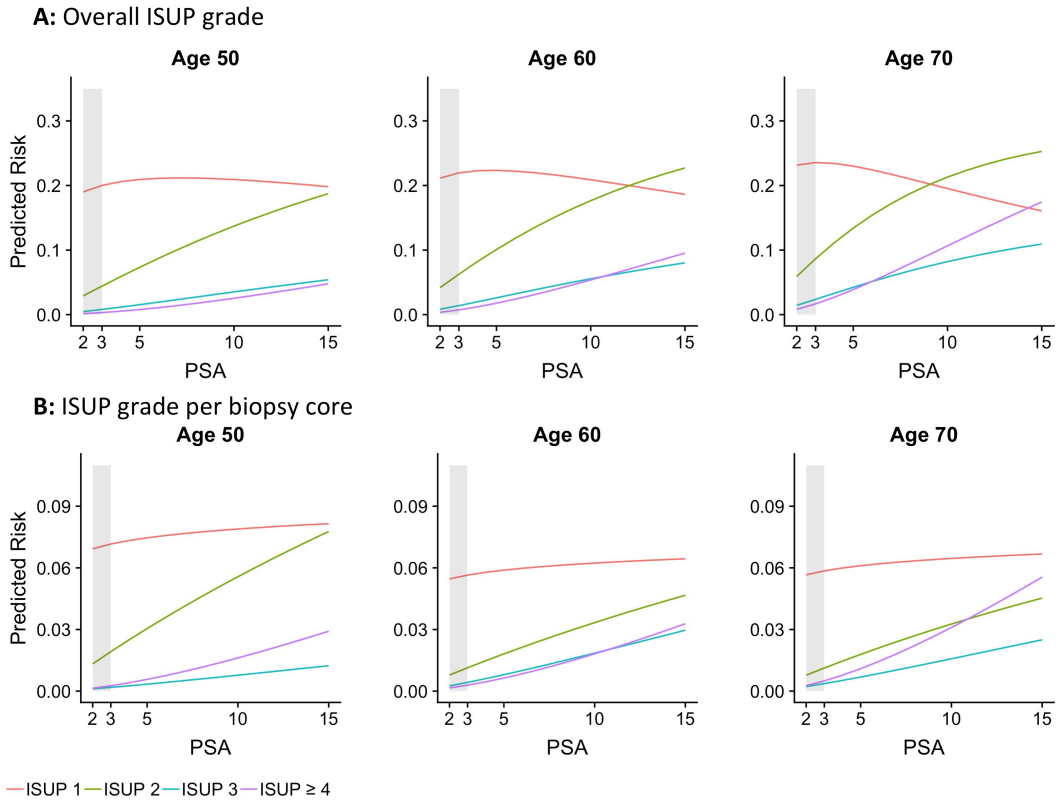
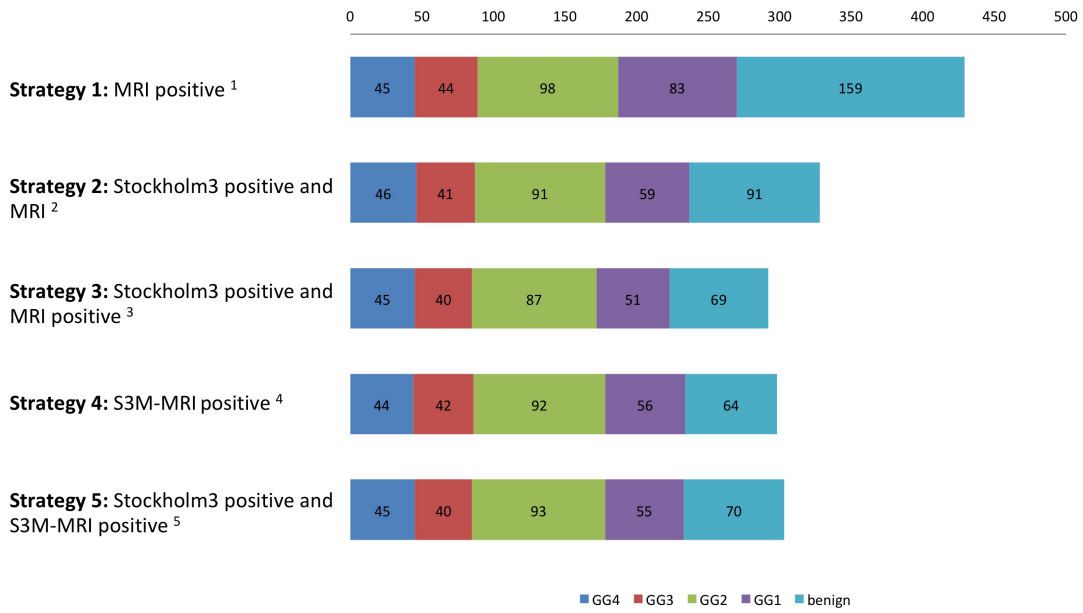


Figure 6.2: Predicted risk of different ISUP grade prostate cancer from biopsy by PSA and age. Risks are predicted both on the overall ISUP grade level and the biopsy core level.

Our results showed that the unified S3M-MRI model had superior predictive abilities compared to the Stockholm3 model and the PI-RADS score to predict the risk of ISUP ≥ 2 prostate cancer. The AUC of the S3M-MRI model was significantly higher than of the Stockholm3 test and the PI-RADS model. The unified S3M-MRI model also had a higher net benefit on the decision curve analysis in comparison to the Stockholm3 test and the PI-RADS model for clinically relevant thresholds for biopsy recommendation.

However, when comparing different diagnostic strategies, using the Stockholm3 test first followed by MRI for the men that tested positive on the Stockholm3 (Strategy 3 vs. Strategy 4 in Figure 6.3) resulted in a similar number of ISUP 1 prostate cancers (56 vs. 51) and benign biopsies (64 vs. 69) while detecting similar number of ISUP ≥ 2 cancers (172 vs. 178). Strategy 3 creates a simpler workflow with only one screening test and much fewer MRI scans than using the S3M-MRI model.

In conclusion, the predictive abilities of the unified S3M-MRI model were superior to the Stockholm3 model and the PI-RADS score from MRI alone. However, the improvement was small compared to using the Stockholm3 test followed by MRI, and resulted in more MRI scans.



[1] Men with a PI-RADS ≥ 3 in MRI and a systematic and targeted biopsy
[2] Men with a Stockholm3 risk score above the cutoff of 0.10 and a systematic biopsy on MRI negative men and a targeted and systematic biopsy on MRI positive men
[3] Men with a Stockholm3 risk score above the cutoff of 0.10 and then an MRI and a systematic and/or targeted biopsy for men positive on both
[4] Men with a S3M-MRI risk score above the cutoff 0.18 and a systematic and/or targeted biopsy
[5] Men with a Stockholm3 risk score above the cutoff 0.09 and a S3M-MRI risk score above 0.12 and a systematic and/or targeted biopsy

Figure 6.3: Clinical outcomes of five diagnostic strategies using the Stockholm3 test, MRI and the S3M-MRI model among 532 men in the Stockholm3-MRI phase 1 study.

6.4 Study IV

In Study IV we investigated the effect of reclassification of ISUP grade (RG) between prostate biopsy and radical prostatectomy specimens on the predictive performance of prostate cancer risk prediction models. We calculated the AUC for four simulated model scenarios with and without error in the disease status (the outcome variable). Furthermore, we used data from the STHLM3 Radical Prostatectomy Cohort to compute the AUC of the Stockholm3 test to predict significant cancer, defined on biopsy and prostatectomy samples, respectively.

The results from our simulations showed that fitting a model with error in the outcome variable leads to a small decline in true predictive performance, but to a large decline in apparent predictive performance when evaluating the models on data with error. Simulation of scenario 1 with no error in the outcome variable of our model (fitting or evaluation) resulted in the highest AUC. Simulation of scenario 2 with error in the outcome variable in the model fitting, but not in the outcome in the model evaluation resulted in only a small decline in the AUC compared to scenario 1. On the other hand, the simulation of scenario 3 and 4, with error in the outcome variable in

the model evaluation, resulted in a much lower AUC then scenario 1 and 2.

To further illustrate the effect of RG in prostate biopsy, we compared the predictive performance of a prostate cancer risk prediction model (such as the Stockholm3 model) to predict clinically significant cancer defined on biopsy (with RG) and radical prostatectomy samples (without RG). We showed with the simulated scenarios that having error in the outcome variable does not decrease the true predictive performance of our prediction model. Moreover, our results showed that the Stockholm3 model has stronger association with clinically significant prostate cancer defined on radical prostatectomy samples (without error) then on biopsy samples (with error).

In conclusion, our results show that RG affects the true predictive abilities a good prostate cancer risk prediction model only to a small degree and that a model with good discriminatory performance like the Stockholm3 test discriminates better between significant and insignificant prostate cancer defined on radical prostatectomy with fewer errors compared to biopsy with more errors.

Chapter 7

Discussion and Conclusions

7.1 PSA Testing and Risk of Prostate Cancer

From our analysis on the Stockholm PSA and Biopsy registry data in Study I, it was evident that unorganized PSA testing in Stockholm was very common in spite of no recommendation for PSA testing by the Swedish National Board of Health and Welfare [77]. Most recommendations for PSA testing state that it can be beneficial for the men to start PSA testing at age 50 years and preferably men should not be tested after the age of 70 or 75 years for healthy men with long life expectancy because the health benefit of screening after that age would be minimal [15, 16, 78]. Even so, the most tested men in Stockholm are older (70–79 years old) [79] and in Study I we showed that of the men tested annually, 56% were 70–74 years old. The current PSA testing pattern has been an eye opener for health policy makers to accept the fact that unorganized PSA testing is common and executed poorly since no organized screening program is available for prostate cancer on a national level. In my opinion, screening in some form would be beneficial from a public health perspective with better organization, meaning that men with high risk of being diagnosed with prostate cancer are followed up properly and men with low risk or less benefit from screening are not overtested. Improved screening methods are also an important link in the implementation of organized screening, i.e. risk prediction models such as the Stockholm3 test, PCPTRC, PBCG, ERSPC, 4K or PHI together with MRI and targeted biopsies [51, 53, 55, 56, 57, 58].

7.1.1 PSA Testing Intervals

If prostate cancer screening is to be implemented in health care systems in some form, a good screening program needs to be designed. The screening program needs to optimize the benefits and minimize the harms of screening. Study I gave insight into how PSA testing intervals are associated with these benefits and harms of screening and the results can be used as one of the references to plan future personalized screening programs to fit men on an individual basis. We showed that for men with PSA above

1 ng/mL, testing with longer than annual intervals was associated with higher risk of ISUP $\geq$ 2 prostate cancer. However, the benefit of an annual PSA testing interval needs to be balanced against the increased risk of cumulative probability of a negative biopsy, which was threefold with annual vs. triennial testing intervals. Even though the screening program will possibly be implemented using a different screening test than the PSA test (such as other available blood tests or prostate cancer risk calculators), all of these include PSA and therefore our results are likely applicable to aid the design of a screening program using any of those risk tools.

7.1.2 ISUP 1 Prostate Cancer and the PSA Test

ISUP 1 prostate cancer is the most common type of prostate cancer. It seldom becomes metastatic and few men die from the disease. Studies have shown that screening men with PSA increases overdiagnosis and overtreatment of prostate cancer [12] because many healthy men are biopsied using PSA as a screening tool for prostate cancer. In Study II we found that PSA was not associated with the diagnosis of ISUP 1 prostate cancer, neither on an overall level nor on a biopsy core level. This evidence suggests that prostate tissue with ISUP 1 prostate cancer behaves more like benign prostate cells then ISUP $\geq$ 2 prostate cancer cells. Many believe that ISUP 1 prostate cancer should not be labeled as cancer and alternatively it can be diagnosed as a type of pre-stage of prostate cancer, keeping those men under surveillance [40]. As long as we only use PSA as an unorganized screening tool for prostate cancer and biopsying healthy men, the diagnosis of ISUP 1 cancers are likely to continue to substitute around half of all diagnosed prostate cancers.

7.2 MRI and Risk Prediction Models

Recent studies have reported that incorporating MRI results into prostate cancer risk prediction models has improved test characteristics compared to models only based on clinical variables [80, 81, 82]. Our results for the unified S3M-MRI model in Study III validated the results of these previous studies and added to them in two respects: we used biomarker variables and a genetic score in addition to the clinical variables and the multisite design of the study allowed validation of the risk model on entirely independent data. However, these improvements in performance of the unified models have to be weighed against the increase in cost of sending all men with suspicion of prostate cancer to MRI and the increase in complexity of the clinical workflow of adding an extra risk prediction model subsequent to MRI examination. Therefore, we also compared clinical outcomes of five diagnostic strategies, using different combinations of the Stockholm3 test, MRI results and the S3M-MRI model. We found that using the S3M-MRI model resulted in similar number of biopsies and diagnosed ISUP

$\geq$ 2 prostate cancers as using the Stockholm3 test and MRI sequentially. The latter strategy resulted in 35% fewer MRI scans without meaningfully affecting the clinical outcomes, thereby saving costs and simplifying workflows compared to MRI risk prediction models.

A recent study by Siddiqui *et al.* showed that the use of prostate MRI can also lead to more accurate classification of biopsy outcome compared to whole gland pathology after prostatectomy [61]. The predictive ability of targeted biopsy for discriminating between ISUP 1 versus ISUP $\geq$ 2 prostate cancer on prostatectomy samples was greater than that of standard biopsy with an AUC of 0.73 compared to AUC of 0.59 with standard biopsy. Most current prostate cancer risk prediction models are trained on systematic biopsy outcome that are to a large degree (30–60%) reclassified when compared to whole gland pathology on radical prostatectomy samples. Thus, using MRI and targeted biopsy results to develop new prostate cancer risk prediction models with improved classification of ISUP grade can improve the predictive abilities of prostate cancer risk prediction models.

7.3 Risk Tools for Prostate Cancer Diagnosis

Most prostate cancer risk tools have an AUC between 0.60 and 0.85 and they are very dependent on the cohort we use due to selection bias. It is common that AUCs are compared between different cohorts, but because of how dependent they are on the type of patients in the cohort, it is almost impossible to compare and should in fact not be done. If we want to compare different risk prediction models, we need to test them on the same cohort of patients and then present our results as an external validation of the risk tools. Another reason for the AUC being lower on average than in good prediction models in other fields is that the outcome variable (systematic biopsy ISUP grade) is to a large degree reclassified on whole gland pathology samples. Thus the true disease status might be wrong in some cases and it is difficult to develop a perfectly discriminating model. Also, PSA has in most studies been used as a pre-selection test for inclusion. If men with low PSA would be included in the datasets for the models, the AUC could improve since it is easier for the model to discriminate between men with very low and very high PSA.

7.4 Limitations

One of the main limitations as well as one of the main strengths of this thesis is the data sources we have used. In the case of the Stockholm3 test, the data set used to develop and validate the model was very large in a clinical setting (n = 59.149 men) using data from men 50–69 years old, of Northern European descent from Stockholm,

Sweden. The model has shown to perform well when predicting the risk of high-grade prostate cancer for Nordic men, both in the validation cohort of our study and in other continued validation studies performed in Stockholm and Norway. To be able to conclude that the model is generalizable for the total population of older men it is an extremely important next step to validate the model on external data using men of different ethnic background, from different geographical areas in countries that have different lifestyles from Swedish men.

Another limitation to this thesis includes a possible **selection bias** in the data. Tangen *et al.* showed that known risk factors for prostate cancer increased the risk of being biopsied in two large prostate cancer prevention trials that can lead to a possible bias in their associations with prostate cancer risk [83]. Risk factors identified in epidemiological studies may be erroneous and can lead to misdirected study conclusions. Because prostate cancer is highly prevalent in older men and it is usually asymptomatic until metastatic, assumptions can be made on risk factors that possibly are not associated with prostate cancer. Respectively, men with that risk factor are more likely to be screened for prostate cancer and subsequently those who screen positive for that risk factor are more likely to be recommended and undergo biopsy. As a result, the men with the risk factor are more likely to be diagnosed with prostate cancer. To counteract this effect, the men in the Stockholm3 study were randomly invited to participate by date of birth and men with high PSA, high Stockholm3 score or those at high risk for prostate cancer in some other aspect were recommended to undergo biopsy.

One more limitation to this thesis is **measurement bias** in our data. Even though registry data and the data we used from the STHLM3 study and Stockholm3-MRI study are of high quality, classification problems in the biopsy outcome are common, which is the outcome variable in all my four studies. Reclassifying the prostatectomy sample with a higher grade than on the biopsy sample is more common than the opposite [43, 45], and thus the risk evaluation in my thesis might be biased towards a lower risk level than the actual risk of prostate cancer. Also, in general we do not have data on the final outcome (death from prostate cancer) and thus we are almost always working with a proxy endpoint.

7.5 Ethical Concerns

Prediction models in a clinical setting can be a very useful tool, and as developers of such powerful tools, we need to be careful to take ethical concerns seriously. It is important to cover all areas of risk, to use the data we have carefully and validate our models in external settings. These are among the most important aspects of developing prediction models, especially when it includes and affects the health and lives of many

people. We also need to ask the right questions in our models and when asking these questions we need to be careful using sensitive information such as gender, race, age and socioeconomic status. There are many examples of harmful risk prediction models that result in discriminating decisions based on personal features. However, in the case of prediction models within medicine we are hopefully always striving to improve the health of individuals and these discriminating features can be important in our model. As long as we deem the model unharmful and exclusively for the purpose of improving the health of individuals, in my opinion it is the right ethical decision to use discriminating features for the development of our models.

The use of population and health registers for research is governed by Swedish law, and using register data under secrecy is allowed for research if the study has been approved by the regional ethics review board and if the register holder approves the data extraction. All studies in this thesis had an ethical approval from the Ethical Review Board in Stockholm. To ensure that the data is handled in a safe manner, registers hold secure databases and safe IT environments and directly identifying information such as name and personal identification number are removed and replaced with a randomly generated id number.

Chapter 8

Future research

The field of prostate cancer diagnostics is currently moving very fast, and there is large number of important research areas within prostate cancer diagnostics where new studies are warranted (e.g. finding and evaluating new markers such as ctDNA, and the use of AI for ISUP grading). I will briefly describe two future studies that I think are key to the development:

8.1 Personalized Screening for Prostate Cancer

Organized screening for prostate cancer is in my opinion the next step in the health care of middle aged men. Whether it will be using PSA only (currently very common and unorganized) or using new risk tools, it is important to study the effect of different screening intervals such as we did in Study 1. Since there has not been any organized prostate cancer screening, there is no direct evidence of the long term effect of different screening intervals on the diagnosis of prostate cancer or mortality rates of the patients after diagnosis.

To study the effect of different screening intervals using risk prediction models, a follow up study of the STHLM3 study is a very interesting option. Such a study can contribute to the research of personalized screening for prostate cancer as well as provide important information on incidence and mortality rates of prostate cancer using organized screening with a risk prediction model. The grand scale of the study cohort, the use of a blood based risk prediction model and by randomly dividing the men into different screening interval groups we could contribute to the design of an optimal screening program for prostate cancer worldwide.

8.2 External Validation of the Stockholm3 and S3M-MRI

To make the Stockholm3 model and the S3M-MRI model applicable for clinical use, an external validation of the models is necessary and extremely important. There are

many factors that come into play when developing a clinical risk prediction model and thus testing the model in other geographical areas, on men of a different ethnicity than Northern-European descent and possibly even younger and older men is an important next step in the validation of the Stockholm3 and the S3M-MRI prostate cancer risk prediction models.

Acknowledgements

There are many people that I would like to thank for their contributions to this thesis, and for their support and encouragement during my years as a PhD student. The most valuable experience and knowledge I have gained in these years comes from all the fantastic people around me at work and away from work.

I would like to express my sincere gratitude to the following:

Martin Eklund, my main supervisor, for sharing his knowledge with me and making me a better scientist, for his willingness to explain and with his kind manner I have learned so much from him, for being ambitious and encouraging me to do good science, for always striving for improvement and for helping me get back on my feet when I got papers rejected and was (really!) disappointed, and last but not least for being a good friend through these four years, I have really enjoyed our conversations about everything and nothing. For all this I am truly grateful and I have really enjoyed being a PhD student under your supervision.

Laufey Tryggvadóttir, my co-supervisor, for being my true spiritual mentor since the day I met her, for choosing me to do this journey to begin with and encouraging me to move to Sweden but at the same time prioritize my family since they will always be the most important part of my life, for teaching me a better balance between family and work, for introducing me to the world of a calm mind and teaching me that we have to exercise the mind as well as the body to be happy, for all this I am truly thankful.

Tobias Nordström, my co-supervisor, for helping me understand prostate cancer and always answering my emails when I need help, for letting me sit in during an operation to see some cool robots and for good scientific conversations on the phone when the emails got too long, it has been a pleasure.

Markus Aly, my co-supervisor, for introducing me to the clinic and sharing his knowledge with me, for letting me touch a real prostate (seriously!) after watching an operation, for good conversations over dinner and after dinner at meetings and for saying yes to teaching me how to cross-country ski (even though we never went out!), let me know when you want some more hákarl!

Mark Clements, my co-supervisor, for welcoming me to Sweden when I first got here and helping me find a home for me and my family, for getting in a big car with me on a road-trip in Iceland, for all the pleasant conversations over coffee, for always staying true to yourself and your beliefs, for having your door open when I need advice, you have a kind heart and are truly a good scientist with all your nerdy programming stuff!

Henrik Grönberg, for being an inspiration as a scientist and creating such a wonderful research environment with a group full of talented people.

Paul Dickman, my mentor, for that one lunch we had as mentor and student! Despite our lack of time spent together you are truly a good mentor and I admire your ambition for good teaching and how much you value that part of the academic work, I could not agree with you more.

My roommates and friends: **Elisabeth Dahlquist, Andreas Karlsson, Bénédicte Delcoigne, Shuang Hao** and **Rikard Strandberg** for being such great roommates and friends. I will remember all the fun conversations with you guys and you all made my time at MEB so colorful and fun.

My friends at MEB: **Therese Andersson, Caroline Weibull, Cecilia Radkiewicz, Maya Alsheh Ali, Kat Bokenberger, Henrik Olsson, Peter Ström** and **Gabriel Isheden**, I have really enjoyed your company and I am grateful for all the laughs we had together over the last years.

Alessio Crippa, Kimmo Kartasalo, Mortezavi Ashkan, Andrea Discacciati, Jan Chandra, Venkatesh Chellappa, Anna Lantz, Mattias Rantalainen, Johan Lindberg, Ola Steinberg, Fredrik Jäderling and **Bram De Laere** in the Prostate Cancer Group, for the good times we have shared at conferences and the research boarding school and for the scientific support, advice and cooperation.

Therese Andersson and **Andrea Discacciati** for reading over my kappa, I am really grateful for your help and input.

Past and present PhD students at MEB for all the nice conversations over lunch or after-work drinks: **Hannah Bower, Linda Abrahamsson, Isabella Ekheden, Malin Ericsson, Alessandra Grotta, Frida Lundberg, Anna Plym, Laura Ghirardi, Elisa Longinetti, Marco Trevisan, Andreas Jangmo, Emilio Ugalde, Xingrong Liu, Zheng Ning, Johan Zetterquist, Daniela Mariosa, Wenjiang Deng** and **Tor-Arne Hegvik**.

Other great colleagues at MEB: **Cecilia Lundholm, Alex Ploner, Keith Humphreys, Anna Johansson, Erin Gabriel, Flaminia Chiesa, Julien Bryois, Sophie Debonneville, Juni Palmgren, Henric Winell, Marie Reilly, Sven Sandin, Arvid Sjölander,**

Agnieszka Szwajda, Nghia Trung Vu, Yudi Pawitan, Dylan Williams and **Robert Karlsson**. You have made my time at MEB so great!

Marie Jansson, Gunilla Sonnenbring and **Camilla Ahlqvist**, thank you for looking after me and making sure that all forms are filled out and papers are scanned for the very distracted PhD student!

Sandra Eloranta and **Karin Ekström Smedby**, for giving me future opportunities within the field of cancer research. I look forward to our cooperation.

My parents, **Sigríður** and **Páll** for always supporting me in the things I choose to do in life and being the best parents I could have. I really appreciate everything you have done for me through the years and you have always been an inspiration to me. Thank you mom for reading over the kappa with me, it has been a great help!

My sisters, **Ólöf** and **Sigurrós**, for being my best friends, for being the best sisters in the world and for always supporting me in everything I do.

My childhood friends, **Anna Helga, Björg Rún, Elísabet, Guðrún Erla, Katrín Ósk** and **Kristín** - you are the best friends one could ever have and I am eternally grateful for your friendship.

My good friends **Ella, Konni, Hildur** and **Hjörtur**, for the crappy dinner parties, the nice trips and all the support during the last four years. I am really happy that Sweden brought us together!

Arnaldur Hilmisson, my husband, for giving me the opportunity to get a Phd degree, because without you I would not have been able to, for your love and patience and teaching me so many things in life, for giving me the three rascals we are trying to raise together and for always doing your best at everything you do, for helping me have distractions in life outside work and introducing me to our many hobbies that have given me so much joy, you are a true inspiration to me and a solid rock in my life.
Sigríður Bríet, þú munt alltaf vera ljósið mitt og með gleðinni þinni og ástinni beint inn í hjartað munt þú alltaf lýsa upp mitt líf. Ég elska þig.
Héðinn, með bjarta brosið þitt og hlýja hjartað hjálpar þú mér að finna tilganginn í lífinu og njóta þess að vera mamma ykkar. Ég elska þig.
Hilmir Páll, með gleði, hlýju og ákveðni hönd í hönd átt þú stóran stað í hjarta mínu og ég er svo þakklát fyrir að eiga þig fyrir son. Ég elska þig.

This work was supported by **NIASC** (Nordic Information for Action eScience Center), **Cancerfonden** (the Swedish Cancer Society), **Nordforsk**, **VR** (Vetenskaps Rådet) and **FORTE** (Forskningsrådet för hälsa, arbetsliv och välfärd).

References

[1] F Lam, M Colombet, L Mery, M Pineros, A Znaor, I Soerjomataram, F Bray, J Ferlay, and M Ervik. Global cancer observatory: Cancer today. http://gco.iarc.fr/today. Lyon, France: International Agency for Research on Cancer. Online accessed: 19.02.2019.

[2] J Ferlay, I Soerjomataram, and M Ervik. Cancer incidence and mortality worldwide: IARC cancerbase no. 11. http://globocan.iarc.fr. Online accessed: 01.03.2019.

[3] David Petterson, Niklas Toorell, and Lars Holmberg. Statistics on cancer incidence 2017. http://socialstyrelsen.se. Online accessed: 20.02.2019.

[4] G Engholm, G Ferlay, and N Christensen. NORDCAN: Cancer Incidence, Mortality, Prevalence and Survival in the Nordic Countries, Version 7.3. http://ancr.nu. Online accessed: 01.03.2019.

[5] Rune Kvåle, Anssi Auvinen, Hans-Olov Adami, Åsa Klint, Eivor Hernes, Bjørn Møller, Eero Pukkala, Hans H Storm, Laufey Tryggvadottir, Steinar Tretli, et al. Interpreting trends in prostate cancer incidence and mortality in the five Nordic countries. *Journal of the National Cancer Institute*, 99(24):1881–1887, 2007.

[6] Fritz H Schröder, Jonas Hugosson, Monique J Roobol, Teuvo LJ Tammela, Stefano Ciatto, Vera Nelen, Maciej Kwiatkowski, Marcos Lujan, Hans Lilja, Marco Zappa, et al. Screening and prostate-cancer mortality in a randomized European study. *New England Journal of Medicine*, 360(13):1320–1328, 2009.

[7] Pär Stattin, Sigrid Carlsson, Benny Holmström, Andrew Vickers, Jonas Hugosson, Hans Lilja, and Håkan Jonsson. Prostate cancer mortality in areas with high and low prostate cancer incidence. *Journal of the National Cancer Institute*, 106(3):dju007, 2014.

[8] Ruth Etzioni, Alex Tsodikov, Angela Mariotto, Aniko Szabo, Seth Falcon, Jake Wegelin, Dante Ditommaso, Kent Karnofski, Roman Gulati, David F. Penson, and Eric Feuer. Quantifying the role of PSA screening in the US prostate cancer mortality decline. *Cancer Causes and Control*, 19(2):175–181, 2008.

[9] Gerald L Andriole, E David Crawford, Robert L Grubb III, Saundra S Buys, David Chia, Timothy R Church, Mona N Fouad, Edward P Gelmann, Paul A Kvale, Douglas J Reding, et al. Mortality results from a randomized prostate-cancer screening trial. *New England Journal of Medicine*, 360(13):1310–1319, 2009.

[10] Scott E Eggener, Ketan Badani, Daniel A Barocas, Glen W Barrisford, Jed-Sian Cheng, Arnold I Chin, Anthony Corcoran, Jonathan I Epstein, Arvin K George, Gopal N Gupta, et al. Gleason 6 prostate cancer: translating biology into population health. *The Journal of Urology*, 194(3):626–634, 2015.

[11] Hashim Uddin Ahmed, Manit Arya, Alex Freeman, and Mark Emberton. Do low-grade and low-volume prostate cancers bear the hallmarks of malignancy? *The Lancet Oncology*, 13(11):e509–e517, 2012.

[12] Stacy Loeb, Marc A Bjurlin, Joseph Nicholson, Teuvo L Tammela, David F Penson, H Ballentine Carter, Peter Carroll, and Ruth Etzioni. Overdiagnosis and overtreatment of prostate cancer. *European Urology*, 65(6):1046–1055, 2014.

[13] Pim J Van Leeuwen, Monique J Roobol, Ries Kranse, Marco Zappa, Sigrid Carlsson, Meelan Bul, Xiaoye Zhu, Chris H Bangma, Fritz H Schröder, and Jonas Hugosson. Towards an optimal interval for prostate cancer screening. *European Urology*, 61(1):171–176, 2012.

[14] Roman Gulati, John L Gore, and Ruth Etzioni. Comparative effectiveness of alternative prostate-specific antigen-based prostate cancer screening strategies: model estimates of potential benefits and harms. *Annals of Internal Medicine*, 158(3):145–153, 2013.

[15] Durado D Brooks, Andrew Wolf, Robert A Smith, Chiranjeev Dash, and Idris Guessous. Prostate cancer screening 2010: updated recommendations from the American Cancer Society. *Journal of the National Medical Association*, 102(5):423–429, 2010.

[16] H Ballentine Carter, Peter C Albertsen, Michael J Barry, Ruth Etzioni, Stephen J Freedland, Kirsten Lynn Greene, Lars Holmberg, Philip Kantoff, Badrinath R Konety, Mohammad Hassan Murad, David F Penson, and Anthony L Zietman. Early detection of prostate cancer: AUA Guideline. *The Journal of Urology*, 190(2):419–26, 2013.

[17] National Comprehensive Cancer Network. Prostate cancer (version 1.2016). https://www.nccn.org/patients/guidelines/prostate. Online accessed: 23.05.-2019.

[18] N Mottet, J Bellmunt, E Briers, RCN Van den Bergh, M Bolla, NJ Van Casteren, P Cornford, S Culine, S Joniau, T Lam, et al. Guidelines on prostate cancer. *European Urology*, 65(1):124–37, 2014.

[19] Virginia A Moyer. Screening for prostate cancer: U.S. preventive services task force recommendation statement. *Annals of Internal Medicine*, 157(2):120–134, 2012.

[20] Kirsten Bibbins-Domingo, David C Grossman, Susan J Curry, and Roobol MJ. The US preventive services task force 2017 draft recommendation statement on screening for prostate cancer. *Jama*, 317(19):1949, 2017.

[21] Jan Adolfsson, Hans Garmo, Eberhard Varenhorst, Göran Ahlgren, Christer Ahlstrand, Ove Andrén, Anna Bill-Axelson, Ola Bratt, Jan-Erik Damber, Karin Hellström, et al. Clinical characteristics and primary treatment of prostate cancer in Sweden between 1996 and 2005: Data from the national prostate cancer register in Sweden. *Scandinavian Journal of Urology and Nephrology*, 41(6):456–477, 2007.

[22] Kathryn K Hodge, John E McNeal, Martha K Terris, and Thomas A Stamey. Random systematic versus directed ultrasound guided transrectal core biopsies of the prostate. *The Journal of Urology*, 142(1):71–74, 1989.

[23] Joseph C Presti, James J Chang, Vivek Bhargava, and Katsuto Shinohara. The optimal systematic prostate biopsy scheme should include 8 rather than 6 biopsies: results of a prospective clinical trial. *The Journal of Urology*, 163(1):163–167, 2000.

[24] Joseph C Presti Jr. Prostate biopsy strategies. *Nature Reviews Urology*, 4(9):505, 2007.

[25] Atsushi Ochiai and R Joseph Babaian. Update on prostate biopsy technique. *Current Opinion in Urology*, 14(3):157–162, 2004.

[26] Jonathan I Epstein, William C Allsbrook Jr, Mahul B Amin, Lars L Egevad, ISUP Grading Committee, et al. The 2005 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *The American Journal of Surgical Pathology*, 29(9):1228–1242, 2005.

[27] Jonathan I Epstein. An update of the gleason grading system. *The Journal of urology*, 183(2):433–440, 2010.

[28] Jonathan I Epstein, Lars Egevad, Mahul B Amin, Brett Delahunt, John R Srigley, and Peter A Humphrey. The 2014 international society of urological pathology

(isup) consensus conference on gleason grading of prostatic carcinoma. *The American journal of surgical pathology*, 40(2):244–252, 2016.

[29] National Cancer Institute U.S. National Cancer Institutes of Health. SEER training modules, Gleason grade. https://training.seer.cancer.gov. Online accessed: 19.03.2019.

[30] Jonathan I Epstein, Alan W Partin, Jurgita Sauvageot, and Patrick C Walsh. Prediction of progression following radical prostatectomy: a multivariate analysis of 721 men with long-term follow-up. *The American Journal of Surgical Pathology*, 20(3):286–292, 1996.

[31] Garth A Green, Alexandra L Hanlon, Tahseen Al-Saleem, and Gerald E Hanks. A gleason score of 7 predicts a worse outcome for prostate carcinoma patients treated with radiotherapy. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 83(5):971–976, 1998.

[32] Lars Egevad, T Granfors, L Karlberg, A Bergh, and Per Stattin. Prognostic value of the Gleason score in prostate cancer. *BJU International*, 89(6):538–542, 2002.

[33] Patricia A Ganz, John M Barry, Wylie Burke, Nananda F Col, and Phaedra S Corso. Annals of Internal Medicine conference National Institutes of Health State-of-the-Science conference : Role of active surveillance in the management of men with localized prostate cancer. *Annals of Internal Medicine*, 156(8), 2012.

[34] Alena Böker, Markus A Kuczyk, Mario W Kramer, Axel S Merseburger, Katharina Krüger, Florian Imkamp, and Christoph A von Klot. True Incidence of Gleason 6 Pathology in Patients with Metastatic Castration Resistant Prostate Cancer (mCRPC). *Advances in Therapy*, 34(1):171–179, 2017.

[35] Hima Bindu Musunuru, Toshihiro Yamamoto, Laurence Klotz, Gabriella Ghanem, Alexandre Mamedov, Peraka Sethukavalan, Vibhuti Jethava, Suneil Jain, Liying Zhang, Danny Vesprini, and Andrew Loblaw. Active surveillance for intermediate risk prostate cancer: survival outcomes in the Sunnybrook experience. *The Journal of Urology*, 196(6):1651–1658, 2016.

[36] Hillary M Ross, Oleksandr N Kryvenko, Janet E Cowan, Jeffry P Simko, Thomas M Wheeler, and Jonathan I Epstein. Do adenocarcinomas of the prostate with Gleason score (GS) 6 have the potential to metastasize to lymph nodes? *The American Journal of Surgical Pathology*, 36(9):1346, 2012.

[37] Sam Watts, Geraldine Leydon, Brian Birch, Philip Prescott, Lily Lai, Susan Eardley, and George Lewith. Depression and anxiety in prostate cancer: a systematic review and meta-analysis of prevalence rates. *BMJ Open*, 4(3):e003901, 2014.

[38] LE Carlson, Maureen Angen, Jodi Cullum, Eillen Goodey, Jan Koopmans, Lisa Lamont, JH MacRae, M Martin, Guy Pelletier, John Robinson, et al. High levels of untreated distress and fatigue in cancer patients. *British Journal of Cancer*, 90(12):2297, 2004.

[39] Saiful Miah, Hashim Ahmed, Alex Freeman, and Mark Emberton. OPINION: Does true Gleason pattern 3 merit its cancer descriptor? *Nature Reviews Urology*, 13(9):541–548, 2016.

[40] H. Ballentine Carter, Alan W. Partin, Patrick C. Walsh, Bruce J. Trock, Robert W. Veltri, William G. Nelson, Donald S. Coffey, Eric A. Singer, and Jonathan I. Epstein. Gleason score 6 aadenocarcinoma: Should it be labeled as cancer? *Journal of Clinical Oncology*, 30(35):4294–4296, 2012.

[41] David D Ørsted, Børge G Nordestgaard, Gorm B Jensen, Peter Schnohr, and Stig E Bojesen. Prostate-specific antigen and long-term prediction of prostate cancer incidence and mortality in the general population. *European urology*, 61(5):865–874, 2012.

[42] Regionala Cancercentrum i Samverkan. Prostatacancer. *Nationellt vårdprogram*, 2017.

[43] Michael Goodman, Kevin C Ward, Adeboye O Osunkoya, Milton W Datta, Daniel Luthringer, Andrew N Young, Katerina Marks, Vaunita Cohen, Jan C Kennedy, Michael J Haber, et al. Frequency and determinants of disagreement and error in gleason scores: A population-based study of prostate cancer. *The Prostate*, 72(13):1389–1398, 2012.

[44] Burkhard Helpap and Lars Egevad. The significance of modified gleason grading of prostatic carcinoma in biopsy and radical prostatectomy specimens. *Virchows Archiv*, 449(6):622–627, 2006.

[45] Jonathan I Epstein, Zhaoyong Feng, Bruce J Trock, and Phillip M Pierorazio. Upgrading and downgrading of prostate cancer from biopsy to radical prostatectomy: incidence and predictive factors using the modified gleason grading system and factoring in tertiary grades. *European urology*, 61(5):1019–1024, 2012.

[46] Daniela Danneman, Linda Drevin, Brett Delahunt, Hemamali Samaratunga, David Robinson, Ola Bratt, Stacy Loeb, Pär Stattin, and Lars Egevad. Accuracy of prostate biopsies for predicting Gleason score in radical prostatectomy specimens: nationwide trends 2000–2012. *BJU International*, 119(1):50–56, 2017.

[47] Chris Morash, Rovena Tey, Chika Agbassi, Laurence Klotz, Tom McGowan, John Srigley, and Andrew Evans. Active surveillance for the management of localized

prostate cancer: guideline recommendations. *Canadian Urological Association Journal*, 9(5-6):171, 2015.

[48] American Cancer Society medical and editorial content team. Radiation therapy for prostate cancer. https://www.cancer.org/cancer/prostate-cancer/treating/radiation-therapy.html. Onlince accessed: 28.05.2019.

[49] Hashim U Ahmed, Ahmed El-Shater Bosaily, Louise C Brown, Rhian Gabe, Richard Kaplan, Mahesh K Parmar, Yolanda Collaco-Moraes, Katie Ward, Richard G Hindley, Alex Freeman, et al. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *The Lancet*, 389(10071):815–822, 2017.

[50] Ian M Thompson, Donna K Pauler, Phyllis J Goodman, Catherine M Tangen, M Scott Lucia, Howard L Parnes, Lori M Minasian, Leslie G Ford, Scott M Lippman, E David Crawford, et al. Prevalence of prostate cancer among men with a prostate-specific antigen level 4.0 ng per milliliter. *New England Journal of Medicine*, 350(22):2239–2246, 2004.

[51] Henrik Grönberg, Jan Adolfsson, Markus Aly, Tobias Nordström, Peter Wiklund, Yvonne Brandberg, James Thompson, Fredrik Wiklund, Johan Lindberg, Mark Clements, Lars Egevad, and Martin Eklund. Prostate cancer screening in men aged 50–69 years (STHLM3): a prospective population-based diagnostic study. *The Lancet Oncology*, 16(16):1667–1676, 2015.

[52] Peter Ström, Tobias Nordström, Henrik Grönberg, and Martin Eklund. The Stockholm3 model for prostate cancer detection: algorithm update, biomarker contribution, and reflex test potential. *European Urology*, 2017.

[53] Ian M Thompson, Donna Pauler Ankerst, Chen Chi, Phyllis J Goodman, Catherine M Tangen, M Scott Lucia, Ziding Feng, Howard L Parnes, and Charles A Coltman Jr. Assessing prostate cancer risk: results from the prostate cancer prevention trial. *Journal of the National Cancer Institute*, 98(8):529–534, 2006.

[54] Donna P Ankerst, Josef Hoefler, Sebastian Bock, Phyllis J Goodman, Andrew Vickers, Javier Hernandez, Lori J Sokoll, Martin G Sanda, John T Wei, Robin J Leach, et al. Prostate cancer prevention trial risk calculator 2.0 for the prediction of low-vs high-grade prostate cancer. *Urology*, 83(6):1362–1368, 2014.

[55] Donna P Ankerst, Johanna Straubinger, Katharina Selig, Lourdes Guerrios, Amanda De Hoedt, Javier Hernandez, Michael A Liss, Robin J Leach, Stephen J Freedland, Michael W Kattan, et al. A contemporary prostate biopsy risk calculator based on multiple heterogeneous cohorts. *European Urology*, 74(2):197–203, 2018.

[56] Monique J Roobol, Heidi A van Vugt, Stacy Loeb, Xiaoye Zhu, Meelan Bul, Chris H Bangma, Arno GLJH van Leenders, Ewout W Steyerberg, and Fritz H Schröder. Prediction of prostate cancer risk: the role of prostate volume and digital rectal examination in the ERSPC risk calculators. *European Urology*, 61(3):577–583, 2012.

[57] Andrew J Vickers, Angel M Cronin, Gunnar Aus, Carl-Gustav Pihl, Charlotte Becker, Kim Pettersson, Peter T Scardino, Jonas Hugosson, and Hans Lilja. A panel of kallikrein markers can reduce unnecessary biopsy for prostate cancer: data from the European Randomized Study of Prostate Cancer Screening in Göteborg, Sweden. *BMC Medicine*, 6:19, 2008.

[58] Jeffrey J Tosoian, Sasha C Druskin, Darian Andreas, Patrick Mullane, Meera Chappidi, Sarah Joo, Kamyar Ghabili, Mufaddal Mamawala, Joseph Agostino, Herbert B Carter, Alan W Partin, Lori J Sokoll, and Ashley E Ross. Prostate Health Index density improves detection of clinically significant prostate cancer. *BJU International*, 2017.

[59] Matthias Röthke, D Blondin, HP Schlemmer, and T Franiel. PI-RADS classification: structured reporting for MRI of the prostate. *Rofo*, 185(3):253–261, 2013.

[60] Morgan R Pokorny, Maarten De Rooij, Earl Duncan, Fritz H Schröder, Robert Parkinson, Jelle O Barentsz, and Leslie C Thompson. Prospective study of diagnostic accuracy comparing prostate cancer detection by transrectal ultrasound: guided biopsy versus magnetic resonance (MR) imaging with subsequent MR-guided biopsy in men without previous prostate biopsies. *European Urology*, 66(2):22–29, 2014.

[61] M Minhaj Siddiqui, Soroush Rais-Bahrami, Baris Turkbey, Arvin K George, Jason Rothwax, Nabeel Shakir, Chinonyerem Okoro, Dima Raskolnikov, Howard L Parnes, W Marston Linehan, et al. Comparison of MR/ultrasound fusion-guided biopsy with ultrasound-guided biopsy for the diagnosis of prostate cancer. *JAMA*, 313(4):390–397, 2015.

[62] Henrik Grönberg, Martin Eklund, Wolfgang Picker, Markus Aly, Fredrik Jäderling, Jan Adolfsson, Martin Landquist, Erik Skaaheim Haug, Peter Ström, Stefan Carlsson, et al. Prostate cancer diagnostics using a combination of the Stockholm3 blood test and multiparametric magnetic resonance imaging. *European Urology*, 74(6):722–728, 2018.

[63] Olivier Rouvière, Philippe Puech, Raphaële Renard-Penna, Michel Claudon, Catherine Roy, Florence Mège-Lechevallier, Myriam Decaussin-Petrucci, Marine Dubreuil-Chambardel, Laurent Magaud, Laurent Remontet, et al. Use of prostate

systematic and targeted biopsy on the basis of multiparametric MRI in biopsy-naive patients (MRI-FIRST): a prospective, multicentre, paired diagnostic study. *The Lancet Oncology*, 20(1):100–109, 2019.

[64] Tobias Nordström, Wolfgang Picker, Markus Aly, Fredrik Jäderling, Jan Adolfsson, Peter Ström, Erik Skaaheim Haug, Martin Eklund, Stefan Carlsson, and Henrik Grönberg. Detection of prostate cancer using a multistep approach with Prostate-specific Antigen, the Stockholm 3 test, and targeted biopsies: The STHLM3 MRI project. *European Urology Focus*, 3(6):526–528, 2017.

[65] Ewout W Steyerberg. *Clinical prediction models: a practical approach to development, validation, and updating*. Springer Science & Business Media, 2008.

[66] Eric Vittinghoff, David V Glidden, Stephen C Shiboski, and Charles E McCulloch. *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*. Springer Science & Business Media, 2011.

[67] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*, volume 1. Springer, 2009.

[68] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[69] Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society for Artificial Intelligence*, 14(771–780):1612, 1999.

[70] Wikipedia contributors. Baron Munchausen – Wikipedia, the free encyclopedia, 2019. Online accessed: 20.06.2019.

[71] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[72] Chris Chatfield. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 158(3):419–444, 1995.

[73] David W Hosmer and Stanley Lemesbow. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics-Theory and Methods*, 9(10):1043–1069, 1980.

[74] Frank Harrell (https://stats.stackexchange.com/users/4253/frank harrell). Evaluating logistic regression and interpretation of Hosmer-Lemeshow Goodness of fit. Cross Validated. URL:https://stats.stackexchange.com/q/207512 (version: 2016-04-15).

[75] Andrew J Vickers and Elena B Elkin. Decision curve analysis: a novel method for evaluating prediction models. *National Institute of Health*, 26(6), 2006.

[76] Ben Van Calster, Laure Wynants, Jan FM Verbeek, Jan Y Verbakel, Evangelia Christodoulou, Andrew J Vickers, Monique J Roobol, and Ewout W Steyerberg. Reporting and interpreting decision curve analysis: a guide for investigators. *European Urology*, 2018.

[77] Swedish National Board for Health and Welfare. Screening för prostatacancer: Rekommendation och bedömningsunderlag, 2019.

[78] David C Grossman, Susan J Curry, Douglas K Owens, Kirsten Bibbins-Domingo, Aaron B Caughey, Karina W Davidson, Chyke A Doubeni, Mark Ebell, John W Epling, Alex R Kemper, et al. Screening for prostate cancer: US Preventive Services Task Force recommendation statement. *JAMA*, 319(18):1901–1913, 2018.

[79] Tobias Nordström, Markus Aly, Mark S Clements, Caroline E Weibull, Jan Adolfsson, and Henrik Grönberg. Prostate-specific antigen (PSA) testing is prevalent and increasing in Stockholm county, Sweden, despite no recommendations for PSA screening: results from a population-based study, 2003–2011. *European Urology*, 63(3):419–425, 2013.

[80] Pim J van Leeuwen, Andrew Hayen, James E Thompson, Daniel Moses, Ron Shnier, Maret Böhm, Magdaline Abuodha, Anne-Maree Haynes, Francis Ting, Jelle Barentsz, et al. A multiparametric magnetic resonance imaging-based risk model to determine the risk of significant prostate cancer prior to biopsy. *BJU International*, 120(6):774–781, 2017.

[81] Jan Philipp Radtke, Manuel Wiesenfarth, Claudia Kesch, Martin T Freitag, Celine D Alt, Kamil Celik, Florian Distler, Wilfried Roth, Kathrin Wieczorek, Christian Stock, et al. Combined clinical parameters and multiparametric magnetic resonance imaging for advanced risk modeling of prostate cancer - patient-tailored risk stratification can reduce unnecessary biopsies. *European Urology*, 72(6):888–896, 2017.

[82] Sherif Mehralivand, Joanna H Shih, Soroush Rais-Bahrami, Aytekin Oto, Sandra Bednarova, Jeffrey W Nix, John V Thomas, Jennifer B Gordetsky, Sonia Gaur, Stephanie A Harmon, et al. A magnetic resonance imaging-based prediction model for prostate biopsy risk stratification. *JAMA Oncology*, 4(5):678–685, 2018.

[83] Catherine M Tangen, Phyllis J Goodman, Cathee Till, Jeannette M Schenk, M Scott Lucia, and Ian M Thompson Jr. Biases in recommendations for and acceptance of prostate biopsy significantly affect assessment of prostate cancer risk factors:

results from two large randomized clinical trials. *Journal of Clinical Oncology*, 34(36):4338, 2016.